

# Hypothesis Testing via Affine Detectors

Anatoli Juditsky \*

Arkadi Nemirovski †

## Abstract

In this paper, we further develop the approach, originating in [13], to “computation-friendly” hypothesis testing via Convex Programming. Most of the existing results on hypothesis testing aim to quantify in a closed analytic form separation between sets of distributions allowing for reliable decision in precisely stated observation models. In contrast to this descriptive (and highly instructive) traditional framework, the approach we promote here can be qualified as operational – the testing routines and their risks are yielded by an efficient computation. All we know in advance is that, under favorable circumstances, specified in [13], the risk of such test, whether high or low, is provably near-optimal under the circumstances. As a compensation for the lack of “explanatory power,” this approach is applicable to a much wider family of observation schemes and hypotheses to be tested than those where “closed form descriptive analysis” is possible.

In the present paper our primary emphasis is on computation: we make a step further in extending the principal tool developed in [13] – testing routines based on affine detectors – to a large variety of testing problems. The price of this development is the loss of blanket near-optimality of the proposed procedures (though it is still preserved in the observation schemes studied in [13], which now become particular cases of the general setting considered here).

## 1 Introduction

This paper can be considered as an extension of [13] where the following simple observation was the starting point of numerous developments:

Imagine that we want to decide on two composite hypotheses about the distribution  $P$  of a random observation  $\omega$  taking values in observation space  $\Omega$ ,  $i$ -th hypothesis stating that  $P \in \mathcal{P}_i$ , where  $\mathcal{P}_i$ ,  $i = 1, 2$ , are given families of probability distributions on  $\Omega$ . Let  $\phi : \Omega \rightarrow \mathbf{R}$  be a *detector*, and let the *risk* of detector  $\phi$  be defined as the smallest  $\epsilon_*$  such that

$$\int_{\Omega} e^{-\phi(\omega)} P(d\omega) \leq \epsilon_* \quad \forall P \in \mathcal{P}_1 \quad \& \quad \int_{\Omega} e^{\phi(\omega)} P(d\omega) \leq \epsilon_* \quad \forall P \in \mathcal{P}_2. \quad (1)$$

Then the test  $\mathcal{T}^K$  which, given  $K$  i.i.d. observations  $\omega_t \sim P \in \mathcal{P}_1 \cup \mathcal{P}_2$ ,  $t = 1, \dots, K$ , deduces that  $P \in \mathcal{P}_1$  when  $\sum_{t=1}^K \phi(\omega_t) \geq 0$ , and that  $P \in \mathcal{P}_2$  otherwise, accepts the true hypothesis with  $P$ -probability at least  $1 - \epsilon_*^K$ .

It was shown in [13] that

---

\*LJK, Université Grenoble Alpes, B.P. 53, 38041 Grenoble Cedex 9, France, [anatoli.juditsky@imag.fr](mailto:anatoli.juditsky@imag.fr)

†Georgia Institute of Technology, Atlanta, Georgia 30332, USA, [nemirovs@isye.gatech.edu](mailto:nemirovs@isye.gatech.edu)

The first author was supported by the CNRS-Mastodons project GARGANTUA, and the LabEx PERSYVAL-Lab (ANR-11-LABX-0025). Research of the second author was supported by NSF grants CMMI-1262063, CCF-1523768.

1. the detector-based tests are “near optimal” – if the above hypotheses can be decided upon by a single-observation test  $\mathcal{T}^*$  with risk  $\delta < 1/2$ , there exists a detector  $\phi$  with “comparable” risk  $\epsilon_\star = 2\sqrt{\delta(1-\delta)}$ .

Note that while the risk  $2\sqrt{\delta(1-\delta)}$  seems to be much larger than  $\delta$ , especially for small  $\delta$ , we can “compensate” for risk deterioration passing from the single-observation test  $\mathcal{T}^1$  associated with the detector  $\phi$  to the test  $\mathcal{T}^K$  based on the same detector and using  $K$  observations. The risk of the test  $\mathcal{T}^K$ , by the above, is upper-bounded by  $\epsilon_\star^K = [2\sqrt{\delta(1-\delta)}]^K$  and thus is not worse than the risk  $\delta$  of the “ideal” single-observation test already for a “quite moderate” value  $\lfloor \frac{2}{1-\ln(4(1-\delta))/\ln(1/\delta)} \rfloor$  of  $K$ .

2. There are “good,” in certain precise sense, parametric families of distributions, primarily,
  - Gaussian  $\mathcal{N}(\mu, I_d)$  distributions on  $\Omega = \mathbf{R}^d$ ,
  - Poisson distributions with parameters  $\mu \in \mathbf{R}_+^d$  on  $\Omega = \mathbf{Z}^d$ ; the corresponding random variables have  $d$  independent entries,  $j$ -th of them being Poisson with parameter  $\mu_j$ ,
  - Discrete distributions on  $\{1, \dots, d\}$ , the parameter  $\mu$  of a distribution being the vector of probabilities to take value  $j = 1, \dots, d$ ,

for which the optimal (with the minimal risk, and thus – near-optimal by 1) detectors can be found efficiently, provided that  $\mathcal{P}_i$ ,  $i = 1, 2$ , are *convex hypotheses*, meaning that they are cut off the family of distributions in question by restricting the distribution’s parameter  $\mu$  to reside in a convex domain. <sup>1</sup>

On a closer inspection, the “common denominator” of Gaussian, Poisson and Discrete families of distributions is that in all these cases the minimal risk detector for a pair of convex hypotheses is *affine*,<sup>2</sup> and the results of [13] in the case of deciding on a pair of convex hypotheses stemming from a *good family of distributions* sum up to the following:

- A) the best – with the smallest possible risk – *affine* detector, same as its risk, can be efficiently computed;
- B) the smallest risk *affine* detector from A) is the best, in terms of risk, detector available under the circumstances, so that the associated test is near-optimal.

Note that as far as practical applications of the above approach are concerned, one “can survive” without B) (near-optimality of the constructed detectors), while A) *is a must*. In this paper, we focus on families of distributions obeying A); this class turns out to be incomparably larger than what was defined as “good” in [13]. In particular, it includes nonparametric families of distributions. Staying within this much broader class, we still are able to construct in a computationally efficient way the best affine detectors for a pair of “convex”, in certain precise sense, hypotheses, along with valid upper bounds on the risks of the detectors. What we, in general, *cannot* claim anymore, is that the tests associated with the above detectors are near-optimal. This being said, we believe that investigating possibilities for building tests and quantifying their performance in a computationally friendly manner is of value even when we cannot provably guarantee near-optimality of these tests.

<sup>1</sup>In retrospect, these results can be seen as a development of the line of research initiated by the pioneering works of H. Chernoff [9], C. Kraft [16], and L. Le Cam [17], further developed in [1, 2, 3, 4, 7, 8] among many others (see also references in [15]).

<sup>2</sup>affinity of a detector makes sense only when  $\Omega$  can be naturally identified with a subset of some  $\mathbf{R}^d$ . This indeed is the case for Gaussian and Poisson distributions; to make it the case for discrete distributions on  $\{1, \dots, d\}$ , it suffices to encode  $j \leq d$  by  $j$ -th basic orth in  $\mathbf{R}^d$ , thus making  $\Omega$  the set of basic orths in  $\mathbf{R}^d$ . With this encoding, every real valued function on  $\{1, \dots, d\}$  becomes affine.

The paper is organized as follows. The families of distributions well suited for constructing affine detectors in a computationally friendly fashion are introduced and investigated in section 2. In particular, we develop a kind of fully algorithmic “calculus” of these families. This calculus demonstrates that the families of probability distributions covered by our approach are much more common commodity than “good observation schemes” as defined in [13]. In section 3 we explain how to build within our framework tests for pairs (and larger tuples) of hypotheses and how to quantify performance of these tests in a computationally efficient fashion. Aside of general results of this type, we work out in detail the case where the family of distributions giving rise to “convex hypotheses” to be tested is comprised of sub-Gaussian distributions (section 3.2.3). In section 4 we discuss an application to the now-classical statistical problem – aggregation of estimators – and show how the results of [12] can be extended to the general situation considered here. Finally, in section 5 we show how our framework can be extended in the Gaussian case to include quadratic detectors. To streamline the presentation, all proofs exceeding few lines are collected in the appendix.

## 2 Setup

Let us fix *observation space*  $\Omega = \mathbf{R}^d$ , and let  $\mathcal{P}_j$ ,  $1 \leq j \leq J$ , be given families of Borel probability distributions on  $\Omega$ . Put broadly, our goal is, given a random observation  $\omega \sim P$ , where  $P \in \bigcup_{j \leq J} \mathcal{P}_j$ , to decide upon the hypotheses  $H_j : P \in \mathcal{P}_j$ ,  $j = 1, \dots, J$ . We intend to address the above goal in the case when the families  $\mathcal{P}_j$  are *simple* – they are comprised of distributions for which moment-generating functions admit an explicit upper bound.

### 2.1 Regular and simple probability distributions

Let

- $\mathcal{H}$  be a nonempty closed convex set in  $\Omega = \mathbf{R}^d$  symmetric w.r.t. the origin,
- $\mathcal{M}$  be a closed convex set in some  $\mathbf{R}^n$ ,
- $\Phi(h; \mu) : \mathcal{H} \times \mathcal{M} \rightarrow \mathbf{R}$  be a continuous function convex in  $h \in \mathcal{H}$  and concave in  $\mu \in \mathcal{M}$ .

We refer to  $\mathcal{H}, \mathcal{M}, \Phi(\cdot, \cdot)$  satisfying the above restrictions as to *regular data*. Regular data  $\mathcal{H}, \mathcal{M}, \Phi(\cdot, \cdot)$  define a family

$$\mathcal{R} = \mathcal{R}[\mathcal{H}, \mathcal{M}, \Phi]$$

of Borel probability distributions  $P$  on  $\Omega$  such that

$$\forall h \in \mathcal{H} \exists \mu \in \mathcal{M} : \ln \left( \int_{\Omega} \exp\{h^T \omega\} P(d\omega) \right) \leq \Phi(h; \mu). \quad (2)$$

We say that distributions satisfying (2) are *regular*, and, given regular data  $\mathcal{H}, \mathcal{M}, \Phi(\cdot, \cdot)$ , we refer to  $\mathcal{R}[\mathcal{H}, \mathcal{M}, \Phi]$  as to *regular family* of distributions associated with the data  $\mathcal{H}, \mathcal{M}, \Phi$ . The same regular data  $\mathcal{H}, \mathcal{M}, \Phi(\cdot, \cdot)$  define a smaller family

$$\mathcal{S} = \mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$$

of Borel probability distributions  $P$  on  $\Omega$  such that

$$\exists \mu \in \mathcal{M} : \forall h \in \mathcal{H} : \ln \left( \int_{\Omega} \exp\{h^T \omega\} P(d\omega) \right) \leq \Phi(h; \mu). \quad (3)$$

We say that distributions satisfying (3) are *simple*. Given regular data  $\mathcal{H}, \mathcal{M}, \Phi(\cdot, \cdot)$ , we refer to  $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$  as to *simple family* of distributions associated with the data  $\mathcal{H}, \mathcal{M}, \Phi$ .

Recall that the starting point of our study is a “plausibly good” detector-based test which, given two families  $\mathcal{P}_1$  and  $\mathcal{P}_2$  of distribution with common observation space, and independent observations  $\omega_1, \dots, \omega_t$  drawn from a distribution  $P \in \mathcal{P}_1 \cup \mathcal{P}_2$ , decides whether  $P \in \mathcal{P}_1$  or  $P \in \mathcal{P}_2$ . Our interest in regular/simple families of distributions stems from the fact that when the families  $\mathcal{P}_1$  and  $\mathcal{P}_2$  are of this type, building such test reduces to solving a convex-concave game and thus can be carried on in a computationally efficient manner. We postpone the related construction and analysis to section 3, and continue with presenting some basic examples of simple and regular families of distributions and a simple “calculus” of these families.

## 2.2 Basic examples of simple families of probability distributions

### 2.2.1 Sub-Gaussian distributions

Let  $\mathcal{H} = \Omega = \mathbf{R}^d$ ,  $\mathcal{M}$  be a closed convex subset of the set  $\mathcal{G}_d = \{\mu = (\theta, \Theta) : \theta \in \mathbf{R}^d, \Theta \in \mathbf{S}_+^d\}$ , where  $\mathbf{S}_+^d$  is cone of positive semidefinite matrices in the space  $\mathbf{S}^d$  of symmetric  $d \times d$  matrices, and let

$$\Phi(h; \theta, \Theta) = \theta^T h + \frac{1}{2} h^T \Theta h.$$

In this case,  $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$  contains all sub-Gaussian distributions  $P$  on  $\mathbf{R}^d$  with sub-Gaussianity parameters from  $\mathcal{M}$ , that is, all Borel probability distributions  $P$  on  $\Omega$  admitting upper bound

$$\mathbf{E}_{\omega \sim P} \{\exp\{h^T \omega\}\} \leq \exp\{\theta^T h + \frac{1}{2} h^T \Theta h\} \quad \forall h \in \mathbf{R}^d \quad (4)$$

on the moment-generating function, with parameters  $(\theta, \Theta)$  of the bound belonging to  $\mathcal{M}$ . In particular,  $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$  contains all Gaussian distributions  $\mathcal{N}(\theta, \Theta)$  with  $(\theta, \Theta) \in \mathcal{M}$ .

### 2.2.2 Poisson distributions

Let  $\mathcal{H} = \Omega = \mathbf{R}^d$ , let  $\mathcal{M}$  be a closed convex subset of  $d$ -dimensional nonnegative orthant  $\mathbf{R}_+^d$ , and let

$$\Phi(h = [h_1; \dots; h_d]; \mu = [\mu_1; \dots; \mu_d]) = \sum_{i=1}^d \mu_i [\exp\{h_i\} - 1] : \mathcal{H} \times \mathcal{M} \rightarrow \mathbf{R}.$$

The family  $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$  contains all product-type Poisson distributions  $\text{Poisson}[\mu]$  with vectors  $\mu$  of parameters belonging to  $\mathcal{M}$ ; here  $\text{Poisson}[\mu]$  is the distribution of random  $d$ -dimensional vector with independent of each other entries,  $i$ -th entry being Poisson random variable with parameter  $\mu_i$ .

### 2.2.3 Discrete distributions

Consider a discrete random variable taking values in  $d$ -element set  $\{1, 2, \dots, d\}$ , and let us think of such a variable as of random variable taking values  $e_i$ ,  $i = 1, \dots, d$ , where  $e_i$  are standard basic orths in  $\mathbf{R}^d$ ; probability distribution of such a variable can be identified with a point  $\mu = [\mu_1; \dots; \mu_d]$  from the  $d$ -dimensional probabilistic simplex

$$\Delta_d = \{\nu \in \mathbf{R}_+^d : \sum_{i=1}^d \nu_i = 1\},$$

where  $\mu_i$  is the probability for the variable to take value  $e_i$ . With these identifications, setting  $\mathcal{H} = \mathbf{R}^d$ , specifying  $\mathcal{M}$  as a closed convex subset of  $\Delta_d$  and setting

$$\Phi(h = [h_1; \dots; h_d]; \mu = [\mu_1; \dots; \mu_d]) = \ln \left( \sum_{i=1}^d \mu_i \exp\{h_i\} \right),$$

the family  $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$  contains distributions of all discrete random variables taking values in  $\{1, \dots, d\}$  with probabilities  $\mu_1, \dots, \mu_d$  comprising a vector from  $\mathcal{M}$ .

#### 2.2.4 Distributions with bounded support

Consider the family  $\mathcal{P}[X]$  of Borel probability distributions supported on a closed and bounded convex set  $X \subset \Omega = \mathbf{R}^d$ , and let

$$\phi_X(h) = \max_{x \in X} h^T x$$

be the support function of  $X$ . We have the following result (to be refined in section 2.3.5):

**Proposition 2.1** *For every  $P \in \mathcal{P}[X]$  it holds*

$$\forall h \in \mathbf{R}^d : \ln \left( \int_{\mathbf{R}^d} \exp\{h^T \omega\} P(d\omega) \right) \leq h^T e[P] + \frac{1}{8} [\phi_X(h) + \phi_X(-h)]^2, \quad (5)$$

where  $e[P] = \int_{\mathbf{R}^d} \omega P(d\omega)$  is the expectation of  $P$ , and the right hand side function in (5) is convex. As a result, setting

$$\mathcal{H} = \mathbf{R}^d, \mathcal{M} = X, \Phi(h; \mu) = h^T \mu + \frac{1}{8} [\phi_X(h) + \phi_X(-h)]^2,$$

we get regular data such that  $\mathcal{P}[X] \subset \mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$ .

For proof, see Appendix A.1

### 2.3 Calculus of regular and simple families of probability distributions

Regular and simple families of probability distributions admit “fully algorithmic” calculus, with the main calculus rules as follows.

#### 2.3.1 Direct summation

For  $1 \leq \ell \leq L$ , let regular data  $\mathcal{H}_\ell \subset \Omega_\ell = \mathbf{R}^{d_\ell}$ ,  $\mathcal{M}_\ell \subset \mathbf{R}^{n_\ell}$ ,  $\Phi_\ell(h_\ell; \mu_\ell) : \mathcal{H}_\ell \times \mathcal{M}_\ell \rightarrow \mathbf{R}$  be given. Let us set

$$\begin{aligned} \Omega &= \Omega_1 \times \dots \times \Omega_L = \mathbf{R}^d, \quad d = d_1 + \dots + d_L, \\ \mathcal{H} &= \mathcal{H}_1 \times \dots \times \mathcal{H}_L = \{h = [h^1; \dots; h^L] : h^\ell \in \mathcal{H}_\ell, \ell \leq L\}, \\ \mathcal{M} &= \mathcal{M}_1 \times \dots \times \mathcal{M}_L = \{\mu = [\mu^1; \dots; \mu^L] : \mu^\ell \in \mathcal{M}_\ell, \ell \leq L\} \subset \mathbf{R}^n, \quad n = n_1 + \dots + n_L, \\ \Phi(h = [h^1; \dots; h^L]; \mu = [\mu^1; \dots; \mu^L]) &= \sum_{\ell=1}^L \Phi_\ell(h^\ell; \mu^\ell) : \mathcal{H} \times \mathcal{M} \rightarrow \mathbf{R}. \end{aligned}$$

Then  $\mathcal{H}$  is a symmetric w.r.t. the origin closed convex set in  $\Omega = \mathbf{R}^d$ ,  $\mathcal{M}$  is a nonempty closed convex set in  $\mathbf{R}^n$ ,  $\Phi : \mathcal{H} \times \mathcal{M} \rightarrow \mathbf{R}$  is a continuous convex-concave function, and clearly

- the family  $\mathcal{R}[\mathcal{H}, \mathcal{M}, \Phi]$  contains all product-type distributions  $P = P_1 \times \dots \times P_L$  on  $\Omega = \Omega_1 \times \dots \times \Omega_L$  with  $P_\ell \in \mathcal{R}[\mathcal{H}_\ell, \mathcal{M}_\ell, \Phi_\ell]$ ,  $1 \leq \ell \leq L$ ;
- the family  $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$  contains all product-type distributions  $P = P_1 \times \dots \times P_L$  on  $\Omega = \Omega_1 \times \dots \times \Omega_L$  with  $P_\ell \in \mathcal{S}[\mathcal{H}_\ell, \mathcal{M}_\ell, \Phi_\ell]$ ,  $1 \leq \ell \leq L$ .

### 2.3.2 IID summation

Let  $\Omega = \mathbf{R}^d$  be an observation space,  $(\mathcal{H}, \mathcal{M}, \Phi)$  be regular data on this space, and let  $\lambda = \{\lambda_\ell\}_{\ell=1}^K$  be a collection of reals. We can associate with the outlined entities a new data  $(\mathcal{H}_\lambda, \mathcal{M}, \Phi_\lambda)$  on  $\Omega$  by setting

$$\mathcal{H}_\lambda = \{h \in \Omega : \|\lambda\|_\infty h \in \mathcal{H}\}, \quad \Phi_\lambda(h; \mu) = \sum_{\ell=1}^L \Phi(\lambda_\ell h; \mu) : \mathcal{H}_\lambda \times \mathcal{M} \rightarrow \mathbf{R}.$$

Now, given a probability distribution  $P$  on  $\Omega$ , we can associate with it and with the above  $\lambda$  a new probability distribution  $P^\lambda$  on  $\Omega$  as follows:  $P^\lambda$  is the distribution of  $\sum_\ell \lambda_\ell \omega_\ell$ , where  $\omega_1, \omega_2, \dots, \omega_L$  are drawn, independently of each other, from  $P$ . An immediate observation is that the data  $(\mathcal{H}_\lambda, \mathcal{M}, \Phi_\lambda)$  is regular, and

- whenever a probability distribution  $P$  belongs to  $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$ , the distribution  $P^\lambda$  belongs to  $\mathcal{S}[\mathcal{H}_\lambda, \mathcal{M}, \Phi_\lambda]$ . In particular, when  $\omega \sim P \in \mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$ , then the distribution  $P^L$  of the sum of  $L$  independent copies of  $\omega$  belongs to  $\mathcal{S}[\mathcal{H}, \mathcal{M}, L\Phi]$ .

### 2.3.3 Semi-direct summation

For  $1 \leq \ell \leq L$ , let regular data  $\mathcal{H}_\ell \subset \Omega_\ell = \mathbf{R}^{d_\ell}$ ,  $\mathcal{M}_\ell$ ,  $\Phi_\ell$  be given. To avoid complications, we assume that for every  $\ell$ ,

- $\mathcal{H}_\ell = \Omega_\ell$ ,
- $\mathcal{M}_\ell$  is bounded.

Let also an  $\epsilon > 0$  be given. We assume that  $\epsilon$  is small, namely,  $L\epsilon < 1$ .

Let us aggregate the given regular data into a new one by setting

$$\mathcal{H} = \Omega := \Omega_1 \times \dots \times \Omega_L = \mathbf{R}^d, \quad d = d_1 + \dots + d_L, \quad \mathcal{M} = \mathcal{M}_1 \times \dots \times \mathcal{M}_L,$$

and let us define function  $\Phi(h; \mu) : \Omega^d \times \mathcal{M} \rightarrow \mathbf{R}$  as follows:

$$\begin{aligned} \Phi(h = [h^1; \dots; h^L]; \mu = [\mu^1; \dots; \mu^L]) &= \inf_{\lambda \in \Delta^\epsilon} \sum_{\ell=1}^d \lambda_\ell \Phi_\ell(h^\ell / \lambda_\ell; \mu^\ell), \\ \Delta^\epsilon &= \{\lambda \in \mathbf{R}^d : \lambda_\ell \geq \epsilon \forall \ell \text{ \& } \sum_{\ell=1}^L \lambda_\ell = 1\}. \end{aligned} \tag{6}$$

By evident reasons, the infimum in the description of  $\Phi$  is achieved, and  $\Phi$  is continuous. In addition,  $\Phi$  is convex in  $h \in \mathbf{R}^d$  and concave in  $\mu \in \mathcal{M}$ . Postponing for a moment verification, the consequences are that  $\mathcal{H} = \Omega = \mathbf{R}^d$ ,  $\mathcal{M}$  and  $\Phi$  form a regular data. We claim that

*Whenever  $\omega = [\omega^1; \dots; \omega^L]$  is a Borel random variable taking values in  $\Omega = \mathbf{R}^{d_1} \times \dots \times \mathbf{R}^{d_L}$ , and the marginal distributions  $P_\ell$ ,  $1 \leq \ell \leq L$ , of  $\omega$  belong to the families  $\mathcal{S}[\mathbf{R}^{d_\ell}, \mathcal{M}_\ell, \Phi_\ell]$  for all  $1 \leq \ell \leq L$ , the distribution  $P$  of  $\omega$  belongs to  $\mathcal{S}[\mathbf{R}^d, \mathcal{M}, \Phi]$ .*

Indeed, since  $P_\ell \in \mathcal{S}[\mathbf{R}^{d_\ell}, \mathcal{M}_\ell, \Phi_\ell]$ , there exists  $\hat{\mu}^\ell \in \mathcal{M}_\ell$  such that

$$\ln(\mathbf{E}_{\omega^\ell \sim P_\ell} \{\exp\{g^T \omega^\ell\}\}) \leq \Phi_\ell(g; \hat{\mu}^\ell) \quad \forall g \in \mathbf{R}^{d_\ell}.$$

Let us set  $\hat{\mu} = [\hat{\mu}^1; \dots; \hat{\mu}^L]$ , and let  $h = [h^1; \dots; h^L] \in \Omega$  be given. We can find  $\lambda \in \Delta^\epsilon$  such that

$$\Phi(h; \hat{\mu}) = \sum_{\ell=1}^L \lambda_\ell \Phi_\ell(h^\ell / \lambda_\ell; \hat{\mu}^\ell).$$

Applying Hölder inequality, we get

$$\mathbf{E}_{[\omega^1; \dots; \omega^L] \sim P} \left\{ \exp \left\{ \sum_{\ell} [h^{\ell}]^T \omega^{\ell} \right\} \right\} \leq \prod_{\ell=1}^L \left( \mathbf{E}_{\omega^{\ell} \sim P_{\ell}} \left\{ [h^{\ell}]^T \omega^{\ell} / \lambda_{\ell} \right\} \right)^{\lambda_{\ell}},$$

whence

$$\ln \left( \mathbf{E}_{[\omega^1; \dots; \omega^L] \sim P} \left\{ \exp \left\{ \sum_{\ell} [h^{\ell}]^T \omega^{\ell} \right\} \right\} \right) \leq \sum_{\ell=1}^L \lambda_{\ell} \Phi_{\ell}(h^{\ell} / \lambda_{\ell}; \hat{\mu}^{\ell}) = \Phi(h; \hat{\mu}).$$

We see that

$$\ln \left( \mathbf{E}_{[\omega^1; \dots; \omega^L] \sim P} \left\{ \exp \left\{ \sum_{\ell} [h^{\ell}]^T \omega^{\ell} \right\} \right\} \right) \leq \Phi(h; \hat{\mu}) \quad \forall h \in \mathcal{H} = \mathbf{R}^d,$$

and thus  $P \in \mathcal{S}[\mathbf{R}^d, \mathcal{M}_{\ell}, \Phi_{\ell}]$ , as claimed.

It remains to verify that the function  $\Phi$  defined by (6) indeed is convex in  $h \in \mathbf{R}^d$  and concave in  $\mu \in \mathcal{M}$ . Concavity in  $\mu$  is evident. Further, functions  $\lambda_{\ell} \Phi_{\ell}(h^{\ell} / \lambda_{\ell}; \mu)$  (as perspective transformation of convex functions  $\Phi_{\ell}(\cdot; \mu)$ ) are jointly convex in  $\lambda$  and  $h^{\ell}$ , and so is  $\Psi(\lambda, h; \mu) = \sum_{\ell=1}^L \lambda_{\ell} \Phi_{\ell}(h^{\ell} / \lambda_{\ell}, \mu)$ . Thus  $\Phi(\cdot; \mu)$ , obtained by partial minimization of  $\Psi$  in  $\lambda$ , indeed is convex.

### 2.3.4 Affine image

Let  $\mathcal{H}, \mathcal{M}, \Phi$  be regular data,  $\Omega$  be the embedding space of  $\mathcal{H}$ , and  $x \mapsto Ax + a$  be an affine mapping from  $\Omega$  to  $\bar{\Omega} = \mathbf{R}^{\bar{d}}$ , and let us set

$$\bar{\mathcal{H}} = \{\bar{h} \in \mathbf{R}^{\bar{d}} : A^T \bar{h} \in \mathcal{H}\}, \quad \bar{\mathcal{M}} = \mathcal{M}, \quad \bar{\Phi}(\bar{h}; \mu) = \Phi(A^T \bar{h}; \mu) + a^T \bar{h} : \bar{\mathcal{H}} \times \bar{\mathcal{M}} \rightarrow \mathbf{R}.$$

Note that  $\bar{\mathcal{H}}, \bar{\mathcal{M}}$  and  $\bar{\Phi}$  are regular data. It is immediately seen that

*Whenever the probability distribution of a random variable  $\omega$  belongs to  $\mathcal{R}[\mathcal{H}, \mathcal{M}, \Phi]$  (or belongs to  $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$ ), the distribution  $\bar{P}[P]$  of the random variable  $\bar{\omega} = A\omega + a$  belongs to  $\mathcal{R}[\bar{\mathcal{H}}, \bar{\mathcal{M}}, \bar{\Phi}]$  (respectively, belongs to  $\mathcal{S}[\bar{\mathcal{H}}, \bar{\mathcal{M}}, \bar{\Phi}]$ ).*

### 2.3.5 Incorporating support information

Consider the situation as follows. We are given regular data  $\mathcal{H} \subset \Omega = \mathbf{R}^d, \mathcal{M}, \Phi$  and are interested in the family of distribution  $\mathcal{P}$  known to belong to  $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$ . In addition, we know that all distributions  $P$  from  $\mathcal{P}$  are supported on a given closed convex set  $X \subset \mathbf{R}^d$ . How could we incorporate this domain information to pass from the family  $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$  containing  $\mathcal{P}$  to a smaller family of the same type still containing  $\mathcal{P}$ ? We are about to give an answer in the simplest case of  $\mathcal{H} = \Omega$ . Specifically, denoting by  $\phi_X(\cdot)$  the support function of  $X$  and selecting somehow a closed convex set  $G \subset \mathbf{R}^d$  containing the origin, let us set

$$\hat{\Phi}(h; \mu) = \inf_{g \in G} [\Phi^+(h, g; \mu) := \Phi(h - g; \mu) + \phi_X(g)],$$

where  $\Phi(h; \mu) : \mathbf{R}^d \times \mathcal{M} \rightarrow \mathbf{R}$  is the continuous convex-concave function participating in the original regular data. Assuming that  $\hat{\Phi}$  is real-valued and continuous on the domain  $\mathbf{R}^d \times \mathcal{M}$  (which definitely is the case when  $G$  is a compact set such that  $\phi_X$  is finite and continuous on  $G$ ), note that  $\hat{\Phi}$  is convex-concave on this domain, so that  $\mathbf{R}^d, \mathcal{M}, \hat{\Phi}$  is a regular data. We claim that

*The family  $\mathcal{S}[\mathbf{R}^d, \mathcal{M}, \hat{\Phi}]$  contains  $\mathcal{P}$ , provided the family  $\mathcal{S}[\mathbf{R}^d, \mathcal{M}, \Phi]$  does so, and the first of these two families is smaller than the second one.*



Verification of the claim is immediate. Let  $P \in \mathcal{P}$ , so that for properly selected  $\mu = \mu_P \in \mathcal{M}$  and for all  $e \in \mathbf{R}^d$  it holds

$$\ln \left( \int_{\mathbf{R}^d} \exp\{e^T \omega\} P(d\omega) \right) \leq \Phi(e; \mu_P).$$

Besides this, for every  $g \in G$  we have  $\phi_X(\omega) - g^T \omega \geq 0$  on the support of  $P$ , whence for every  $h \in \mathbf{R}^d$  one has

$$\ln \left( \int_{\mathbf{R}^d} \exp\{h^T \omega\} P(d\omega) \right) \leq \ln \left( \int_{\mathbf{R}^d} \exp\{h^T \omega + \phi_X(g) - g^T \omega\} P(d\omega) \right) \leq \phi_X(g) + \Phi(h - g; \mu_P).$$

Since the resulting inequality holds true for all  $g \in G$ , we get

$$\ln \left( \int_{\mathbf{R}^d} \exp\{h^T \omega\} P(d\omega) \right) \leq \widehat{\Phi}(h; \mu_P) \quad \forall h \in \mathbf{R}^d,$$

implying that  $P \in \mathcal{S}[\mathbf{R}^d, \mathcal{M}, \widehat{\Phi}]$ ; since  $P \in \mathcal{P}$  is arbitrary, the first part of the claim is justified. The inclusion  $\mathcal{S}[\mathbf{R}^d, \mathcal{M}, \widehat{\Phi}] \subset \mathcal{S}[\mathbf{R}^d, \mathcal{M}, \Phi]$  is readily given by the inequality  $\widehat{\Phi} \leq \Phi$ , and the latter is due to  $\widehat{\Phi}(h, \mu) \leq \Phi(h - 0, \mu) + \phi_X(0)$ .

**Illustration: distributions with bounded support revisited.** In section 2.2.4, given a convex compact set  $X \subset \mathbf{R}^d$  with support function  $\phi_X$ , we checked that the data  $\mathcal{H} = \mathbf{R}^d$ ,  $\mathcal{M} = X$ ,  $\Phi(h; \mu) = h^T \mu + \frac{1}{8}[\phi_X(h) + \phi_X(-h)]^2$  are regular and the family  $\mathcal{S}[\mathbf{R}^d, \mathcal{M}, \Phi]$  contains the family  $\mathcal{P}[X]$  of all Borel probability distributions supported on  $X$ . Moreover, for every  $\mu \in \mathcal{M} = X$ , the family  $\mathcal{S}[\mathbf{R}^d, \{\mu\}, \Phi|_{\mathbf{R}^d \times \{\mu\}}]$  contains all supported on  $X$  distributions with the expectations  $e[P] = \mu$ .

Note that  $\Phi(h; e[P])$  describes well the behaviour of the logarithm  $F_P(h) = \ln \left( \int_{\mathbf{R}^d} e^{h^T \omega} P(d\omega) \right)$  of the moment-generating function of  $P \in \mathcal{P}[X]$  when  $h$  is small (indeed,  $F_P(h) = h^T e[P] + O(\|h\|^2)$  as  $h \rightarrow 0$ ), and by far overestimates  $F_P(h)$  when  $h$  is large. Utilizing the above construction, we replace  $\Phi$  with the real-valued, convex-concave and continuous on  $\mathbf{R}^d \times \mathcal{M}$  function

$$\widehat{\Phi}(h; \mu) = \inf_g \left[ (h - g)^T \mu + \frac{1}{8}[\phi_X(h - g) + \phi_X(-h + g)]^2 + \phi_X(g) \right] \leq \Phi(h; \mu).$$

It is easy to see that  $\widehat{\Phi}(\cdot; \cdot)$  still ensures the inclusion  $P \in \mathcal{S}[\mathbf{R}^d, \{e[P]\}, \widehat{\Phi}|_{\mathbf{R}^d \times \{e[P]\}}]$  for every distribution  $P \in \mathcal{P}[X]$  and “reproduces  $F_P(h)$  reasonably well” for both small and large  $h$ . Indeed, since  $F_P(h) \leq \widehat{\Phi}(h; e[P]) \leq \Phi(h; e[P])$ , for small  $h$   $\widehat{\Phi}(h; e[P])$  reproduces  $F_P(h)$  even better than  $\Phi(h; e[P])$ , and we clearly have

$$\widehat{\Phi}(h; \mu) \leq \left[ (h - h)^T \mu + \frac{1}{8}[\phi_X(h - h) + \phi_X(-h + h)]^2 + \phi_X(h) \right] = \phi_X(h) \quad \forall \mu,$$

and  $\phi_X(h)$  is a correct description of  $F_P(h)$  for large  $h$ .

### 3 Affine detectors and hypothesis testing

#### 3.1 Situation

Assume we are given two collections of regular data with common  $\Omega = \mathbf{R}^d$  and  $\mathcal{H}$ , specifically, the collections  $(\mathcal{H}, \mathcal{M}_\chi, \Phi_\chi)$ ,  $\chi = 1, 2$ . We start with a construction of a specific test for a pair of hypotheses  $H_1 : P \in \mathcal{P}_1$ ,  $H_2 : P \in \mathcal{P}_2$ , where

$$\mathcal{P}_\chi = \mathcal{R}[\mathcal{H}, \mathcal{M}_\chi, \Phi_\chi], \quad \chi = 1, 2.$$

When building the test, we impose on the regular data in question the following



**Assumption I:** The regular data  $(\mathcal{H}, \mathcal{M}_\chi, \Phi_\chi)$ ,  $\chi = 1, 2$ , are such that the convex-concave function

$$\Psi(h; \mu_1, \mu_2) = \frac{1}{2} [\Phi_1(-h; \mu_1) + \Phi_2(h; \mu_2)] : \mathcal{H} \times (\mathcal{M}_1 \times \mathcal{M}_2) \rightarrow \mathbf{R} \quad (7)$$

has a saddle point (min in  $h \in \mathcal{H}$ , max in  $(\mu_1, \mu_2) \in \mathcal{M}_1 \times \mathcal{M}_2$ ).

We associate with a saddle point  $(h_*, \mu_1^*, \mu_2^*)$  the following entities:

- the risk

$$\epsilon_* = \exp\{\Psi(h_*; \mu_1^*, \mu_2^*)\}; \quad (8)$$

this quantity is uniquely defined by the saddle point value of  $\Psi$  and thus is independent of how we select a saddle point;

- the detector  $\phi_*(\omega)$  – the affine function of  $\omega \in \mathbf{R}^d$  given by

$$\phi_*(\omega) = h_*^T \omega + a, \quad a = \frac{1}{2} [\Phi_1(-h_*; \mu_1^*) - \Phi_2(h_*; \mu_2^*)]. \quad (9)$$

A simple sufficient condition for existence of a saddle point of (7) is

**Condition A:** The sets  $\mathcal{M}_1$  and  $\mathcal{M}_2$  are compact, and the function

$$\overline{\Phi}(h) = \max_{\mu_1 \in \mathcal{M}_1, \mu_2 \in \mathcal{M}_2} \Phi(h; \mu_1, \mu_2)$$

is coercive on  $\mathcal{H}$ , meaning that  $\overline{\Phi}(h_i) \rightarrow \infty$  along every sequence  $h_i \in \mathcal{H}$  with  $\|h_i\|_2 \rightarrow \infty$  as  $i \rightarrow \infty$ .

Indeed, under Condition A by Sion-Kakutani Theorem [14] it holds

$$\text{SadVal}[\Phi] := \inf_{h \in \mathcal{H}} \underbrace{\max_{\mu_1 \in \mathcal{M}_1, \mu_2 \in \mathcal{M}_2} \Phi(h; \mu_1, \mu_2)}_{\overline{\Phi}(h)} = \sup_{\mu_1 \in \mathcal{M}_1, \mu_2 \in \mathcal{M}_2} \underbrace{\inf_{h \in \mathcal{H}} \Phi(h; \mu_1, \mu_2)}_{\underline{\Phi}(\mu_1, \mu_2)},$$

so that the optimization problems

$$\begin{aligned} (P) : \quad & \text{Opt}(P) = \min_{h \in \mathcal{H}} \overline{\Phi}(h) \\ (D) : \quad & \text{Opt}(D) = \max_{\mu_1 \in \mathcal{M}_1, \mu_2 \in \mathcal{M}_2} \underline{\Phi}(\mu_1, \mu_2) \end{aligned}$$

have equal optimal values. Under Condition A, problem (P) clearly is a problem of minimizing a continuous coercive function over a closed set and as such is solvable; thus,  $\text{Opt}(P) = \text{Opt}(D)$  is a real. Problem (D) clearly is the problem of maximizing over a compact set of an upper semi-continuous (since  $\Phi$  is continuous) function taking real values and, perhaps, value  $-\infty$ , and not identically equal to  $-\infty$  (since  $\text{Opt}(D)$  is a real), and thus (D) is solvable. Thus, (P) and (D) are solvable with common optimal value, and therefore  $\Phi$  has a saddle point.

## 3.2 Pairwise testing regular families of distributions

### 3.2.1 Main observation

An immediate (and crucial!) observation is as follows:

**Proposition 3.1** *In the situation of section 3.1 and under Assumption I, one has*

$$\begin{aligned} \int_{\Omega} \exp\{-\phi_*(\omega)\} P(d\omega) &\leq \epsilon_* \quad \forall P \in \mathcal{P}_1 = \mathcal{R}[\mathcal{H}, \mathcal{M}_1, \Phi_1] \\ \int_{\Omega} \exp\{\phi_*(\omega)\} P(d\omega) &\leq \epsilon_* \quad \forall P \in \mathcal{P}_2 = \mathcal{R}[\mathcal{H}, \mathcal{M}_2, \Phi_2]. \end{aligned} \quad (10)$$

**Proof.** For every  $\mu_1 \in \mathcal{M}_1$ , we have  $\Phi_1(-h_*; \mu_1) \leq \Phi_1(-h_*; \mu_1^*)$ , and for every  $P \in \mathcal{P}_1$ , we have

$$\int_{\Omega} \exp\{-h_*^T \omega\} P(d\omega) \leq \exp\{\Phi_1(-h_*; \mu_1)\}$$

for properly selected  $\mu_1 \in \mathcal{M}_1$ . Thus,

$$\int_{\Omega} \exp\{-h_*^T \omega\} P(d\omega) \leq \exp\{\Phi_1(-h_*; \mu_1^*)\} \quad \forall P \in \mathcal{P}_1,$$

whence also

$$\int_{\Omega} \exp\{-h_*^T \omega - a\} P(d\omega) \leq \exp\{\frac{1}{2} [\Phi_1(-h_*; \mu_1^*) + \Phi_2(h_*, \mu_2^*)]\} = \epsilon_* \quad \forall P \in \mathcal{P}_1.$$

Similarly, for every  $\mu_2 \in \mathcal{M}_2$ , we have  $\Phi_2(h_*; \mu_2) \leq \Phi_2(h_*; \mu_2^*)$ , and for every  $P \in \mathcal{P}_2$ , we have

$$\int_{\Omega} \exp\{h_*^T \omega\} P(d\omega) \leq \exp\{\Phi_2(h_*; \mu_2)\}$$

for properly selected  $\mu_2 \in \mathcal{M}_2$ . Thus,

$$\int_{\Omega} \exp\{h_*^T \omega\} P(d\omega) \leq \exp\{\Phi_2(h_*; \mu_2^*)\} \quad \forall P \in \mathcal{P}_2,$$

whence

$$\int_{\Omega} \exp\{h_*^T \omega + a\} P(d\omega) \leq \exp\{\frac{1}{2} [\Phi_1(-h_*; \mu_1^*) + \Phi_2(h_*, \mu_2^*)]\} = \epsilon_* \quad \forall P \in \mathcal{P}_2. \quad \square$$

### 3.2.2 Testing pairs of hypotheses

**Repeated observation.** Given  $\Omega = \mathbf{R}^d$ , let random observations  $\omega_t \in \Omega$ ,  $t = 1, 2, \dots$ , be generated as follows:

“In the nature” there exists a random sequence  $\zeta_t \in \mathbf{R}^N$ ,  $t = 1, 2, \dots$ , of *driving factors* such that  $\omega_t$  is a deterministic function of  $\zeta^t = (\zeta_1, \dots, \zeta_t)$ ,  $t = 1, 2, \dots$ ;

we refer to this situation as to case of *repeated observation*  $\omega^\infty$ .

Let now  $(\mathcal{H}, \mathcal{M}, \Phi)$  be a regular data with observation space  $\Omega$ . We associate with this data four hypotheses on the stochastic nature of observations  $\omega^\infty = \{\omega_1, \omega_2, \dots\}$ . Denoting by  $P_{|\zeta^{t-1}}$  the conditional,  $\zeta^{t-1}$  being given, distribution of  $\omega_t$ , we say that the (distribution of the) repeated observation  $\omega^\infty$  obeys hypothesis

- $H_{\mathcal{R}}[\mathcal{H}, \mathcal{M}, \Phi]$ , if  $P_{|\zeta^{t-1}} \in \mathcal{R}[\mathcal{H}, \mathcal{M}, \Phi]$  for all  $t$  and all  $\zeta^{t-1}$ ;
- $H_{\mathcal{S}}[\mathcal{H}, \mathcal{M}, \Phi]$ , if  $P_{|\zeta^{t-1}} \in \mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$  for all  $t$  and all  $\zeta^{t-1}$ ;
- $H_{\mathcal{R}i}[\mathcal{H}, \mathcal{M}, \Phi]$ , if  $P_{|\zeta^{t-1}}$  is independent of  $t$  and  $\zeta^{t-1}$  and belongs to  $\mathcal{R}[\mathcal{H}, \mathcal{M}, \Phi]$ ;
- $H_{\mathcal{S}i}[\mathcal{H}, \mathcal{M}, \Phi]$ , if  $P_{|\zeta^{t-1}}$  is independent of  $t$  and  $\zeta^{t-1}$  and belongs to  $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$ .

Note that the last two hypotheses, in contrast to the first two, require from the observations  $\omega_1, \omega_2, \dots$  to be i.i.d. Note also that  $H_{\mathcal{R}}$  is weaker than  $H_{\mathcal{S}}$ ,  $H_{\mathcal{R}i}$  is weaker than  $H_{\mathcal{S}i}$ , and the “non-stationary” hypotheses  $H_{\mathcal{R}}$ ,  $H_{\mathcal{S}}$  are weaker than their respective stationary counterparts  $H_{\mathcal{R}i}$ ,  $H_{\mathcal{S}i}$ .

The tests to be considered in the sequel operate with the initial fragment  $\omega^K = (\omega_1, \dots, \omega_K)$ , of a prescribed length  $K$ , of the repeated observation  $\omega^\infty = \{\omega_1, \omega_2, \dots\}$ . We call  $\omega^K$  *K-repeated observation* and say that (the distribution of)  $\omega^K$  obeys one of the above hypotheses, if  $\omega^K$  is cut off the repeated observation  $\omega^\infty$  distributed according to the hypothesis in question. We can think of  $H_{\mathcal{R}}$ ,  $H_{\mathcal{S}}$ ,  $H_{\mathcal{R}i}$  and  $H_{\mathcal{S}i}$  as hypotheses about the distribution of the  $K$ -repeated observation  $\omega^K$ .

**Pairwise hypothesis testing from repeated observations.** Assume we are given two collections of regular data  $(\mathcal{H}, \mathcal{M}_\chi, \Phi_\chi)$ ,  $\chi = 1, 2$ , with common observation space  $\Omega = \mathbf{R}^d$  and common  $\mathcal{H}$ . Given positive integer  $K$  and  $K$ -repeated observation  $\omega^K = (\omega_1, \dots, \omega_K)$ , we want to decide on the pair of hypotheses  $H_\chi = H_{\mathcal{R}}[\mathcal{H}, \mathcal{M}_\chi, \Phi_\chi]$ ,  $\chi = 1, 2$ , on the distribution of  $\omega^K$ .

Assume that the convex-concave function (7) associated with the pair of regular data in question has a saddle point  $(h_*; \mu_1^*, \mu_2^*)$ , and let  $\phi_*(\cdot)$ ,  $\epsilon_*$  be the induced by this saddle point affine detector and its risk, see (9), (8). Let us set

$$\phi_*^{(K)}(\omega^K) = \sum_{t=1}^K \phi_*(\omega_t).$$

Consider decision rule  $\mathcal{T}_*^K$  for hypotheses  $H_\chi$ ,  $\chi = 1, 2$ , which, given observation  $\omega^K$ ,

- accepts  $H_1$  (and rejects  $H_2$ ) if  $\phi_*^{(K)}(\omega^K) > 0$ ;
- accepts  $H_2$  (and rejects  $H_1$ ) if  $\phi_*^{(K)}(\omega^K) < 0$ ;
- in the case of a tie (when  $\phi_*^{(K)}(\omega^K) = 0$ ) the test, say, accepts  $H_1$  and rejects  $H_2$ .

In what follows, we refer to  $\mathcal{T}_*^K$  as to *test associated with detector  $\phi_*^{(K)}$* .

**Proposition 3.2** *In the situation in question, we have*

$$\begin{aligned} (a) \quad & \mathbf{E}_{\zeta^K} \left\{ \exp\{-\phi_*^{(K)}(\omega^K)\} \right\} \leq \epsilon_*^K, \quad \text{when } \omega^\infty \text{ obeys } H_1; \\ (b) \quad & \mathbf{E}_{\zeta^K} \left\{ \exp\{\phi_*^{(K)}(\omega^K)\} \right\} \leq \epsilon_*^K, \quad \text{when } \omega^\infty \text{ obeys } H_2. \end{aligned} \tag{11}$$

As a result, the test  $\mathcal{T}_*^K$  accepts exactly one of the hypotheses  $H_1$ ,  $H_2$ , and the risk of this test — the maximal, over  $\chi = 1, 2$ , probability not to accept the hypothesis  $H_\chi$  when it is true (i.e., when the  $K$ -repeated observation  $\omega^K$  obeys the hypothesis  $H_{\mathcal{R}}[\mathcal{H}, \mathcal{M}_\chi, \Phi_\chi]$ ) — does not exceed  $\epsilon_*^K$ .

**Proof.** The fact that the test always accepts exactly one of the hypotheses  $H_\chi$ ,  $\chi = 1, 2$ , is evident. Let us denote  $\mathbf{E}_{\zeta^t}$  the expectation w.r.t. the distribution of  $\zeta^t$ , and let  $\mathbf{E}_{\zeta_{t+1}|\zeta^t}$  stand for expectation w.r.t. conditional to  $\zeta^t$  distribution of  $\zeta_{t+1}$ . Assuming that  $H_1$  holds true and invoking the first inequality in (10), we have

$$\begin{aligned} \mathbf{E}_{\zeta_1} \{ \exp\{-\phi_*^{(1)}(\omega_1)\} \} & \leq \epsilon_*, \\ \mathbf{E}_{\zeta_{t+1}} \{ \exp\{-\phi_*^{(t+1)}(\omega^{t+1})\} \} & = \mathbf{E}_{\zeta^t} \left\{ \exp\{-\phi_*^{(t)}(\omega^t)\} \mathbf{E}_{\zeta_{t+1}|\zeta^t} \{ \exp\{-\phi_*(\omega_{t+1})\} \} \right\} \\ & \leq \epsilon_* \mathbf{E}_{\zeta^t} \left\{ \exp\{-\phi_*^{(t)}(\omega^t)\} \right\}, \quad 1 \leq t < K \end{aligned}$$

(we have taken into account that  $\omega_{t+1}$  is a deterministic function of  $\zeta^{t+1}$  and that we are in the case where the conditional to  $\zeta^t$  distribution of  $\omega_{t+1}$  belongs to  $\mathcal{P}_1 = \mathcal{R}[\mathcal{H}, \mathcal{M}_1, \Phi_1]$ ), and we arrive at

(11.a). This inequality clearly implies that the probability to reject  $H_1$  when the hypothesis is true is  $\leq \epsilon_\star^K$  (since  $\phi_\star^{(K)}(\omega^K) \leq 0$  when  $\mathcal{T}_\star^K$  rejects  $H_1$ ). Assuming that  $H_2$  is true and invoking the second inequality in (10), similar reasoning shows that (11.b) holds true, so that the probability to reject  $H_2$  when the hypothesis is true does not exceed  $\epsilon_\star^K$ .  $\square$

### 3.2.3 Illustration: sub-Gaussian and Gaussian cases

For  $\chi = 1, 2$ , let  $U_\chi$  be nonempty closed convex set in  $\mathbf{R}^d$ , and  $\mathcal{U}_\chi$  be a compact convex subset of the interior of the positive semidefinite cone  $\mathbf{S}_+^d$ . We assume that  $U_1$  is compact. Setting

$$\mathcal{H}_\chi = \Omega = \mathbf{R}^d, \mathcal{M}_\chi = U_\chi \times \mathcal{U}_\chi, \Phi_\chi(h; \theta, \Theta) = \theta^T h + \frac{1}{2} h^T \Theta h : \mathcal{H}_\chi \times \mathcal{M}_\chi \rightarrow \mathbf{R}, \chi = 1, 2, \quad (12)$$

we get two collections  $(\mathcal{H}, \mathcal{M}_\chi, \Phi_\chi)$ ,  $\chi = 1, 2$ , of regular data. As we know from section 2.2.1, for  $\chi = 1, 2$ , the families of distributions  $\mathcal{S}[\mathbf{R}^d, \mathcal{M}_\chi, \Phi_\chi]$  contain the families  $\mathcal{SG}[U_\chi, \mathcal{U}_\chi]$  of sub-Gaussian distributions on  $\mathbf{R}^d$  with sub-Gaussianity parameters  $(\theta, \Theta) \in U_\chi \times \mathcal{U}_\chi$  (see (4)), as well as families  $\mathcal{G}[U_\chi, \mathcal{U}_\chi]$  of Gaussian distributions on  $\mathbf{R}^d$  with parameters  $(\theta, \Theta)$  (expectation and covariance matrix) running through  $U_\chi \times \mathcal{U}_\chi$ . Besides this, the pair of regular data in question clearly satisfies Condition A. Consequently, the test  $\mathcal{T}_\star^K$  given by the above construction as applied to the collections of regular data (12) is well defined and allows to decide on hypotheses  $H_\chi = H_{\mathcal{R}}[\mathbf{R}^d, U_\chi, \mathcal{U}_\chi]$ ,  $\chi = 1, 2$ , on the distribution of the  $K$ -repeated observation  $\omega^K$ . The same test can be also used to decide on stricter hypotheses  $H_\chi^G$ ,  $\chi = 1, 2$ , stating that the observations  $\omega_1, \dots, \omega_K$  are i.i.d. and drawn from a Gaussian distribution  $P$  belonging to  $\mathcal{G}[U_\chi, \mathcal{U}_\chi]$ . Our goal now is to process in detail the situation in question and to refine our conclusions on the risk of the test  $\mathcal{T}_\star^1$  when the *Gaussian* hypotheses  $H_\chi^G$  are considered and the situation is *symmetric*, that is, when  $\mathcal{U}_1 = \mathcal{U}_2$ .

Observe, first, that the convex-concave function  $\Psi$  from (7) in the situation under consideration becomes

$$\Psi(h; \theta_1, \Theta_1, \theta_2, \Theta_2) = \frac{1}{2} h^T [\theta_2 - \theta_1] + \frac{1}{4} h^T \Theta_1 h + \frac{1}{4} h^T \Theta_2 h. \quad (13)$$

We are interested in solutions to the saddle point problem – find a saddle point of function (13) –

$$\min_{h \in \mathbf{R}^d} \max_{\substack{\theta_1 \in U_1, \theta_2 \in U_2 \\ \Theta_1 \in \mathcal{U}_1, \Theta_2 \in \mathcal{U}_2}} \Psi(h; \theta_1, \Theta_1, \theta_2, \Theta_2) \quad (14)$$

From the structure of  $\Psi$  and compactness of  $U_1, \mathcal{U}_1, \mathcal{U}_2$ , combined with the fact that  $\mathcal{U}_\chi$ ,  $\chi = 1, 2$ , are comprised of positive definite matrices, it immediately follows that saddle points do exist, and a saddle point  $(h_\star; \theta_1^\star, \Theta_1^\star, \theta_2^\star, \Theta_2^\star)$  satisfies the relations

$$\begin{aligned} (a) \quad & h_\star = [\Theta_1^\star + \Theta_2^\star]^{-1} [\theta_1^\star - \theta_2^\star], \\ (b) \quad & h_\star^T (\theta_1 - \theta_1^\star) \geq 0 \quad \forall \theta_1 \in U_1, \quad h_\star^T (\theta_2^\star - \theta_2) \geq 0 \quad \forall \theta_2 \in U_2, \\ (c) \quad & h_\star^T \Theta_1 h_\star \leq h_\star^T \Theta_1^\star h_\star \quad \forall \Theta_1 \in \mathcal{U}_1, \quad h_\star^T \Theta_2 h_\star \leq h_\star^T \Theta_2^\star h_\star \quad \forall \Theta_2 \in \mathcal{U}_2. \end{aligned} \quad (15)$$

From (15.a) it immediately follows that the affine detector  $\phi_\star(\cdot)$  and risk  $\epsilon_\star$ , as given by (8) and (9), are

$$\begin{aligned} \phi_\star(\omega) &= h_\star^T [\omega - w_\star] + \frac{1}{2} h_\star^T [\Theta_1^\star - \Theta_2^\star] h_\star, \quad w_\star = \frac{1}{2} [\theta_1^\star + \theta_2^\star]; \\ \epsilon_\star &= \exp\left\{-\frac{1}{4} [\theta_1^\star - \theta_2^\star]^T [\Theta_1^\star + \Theta_2^\star]^{-1} [\theta_1^\star - \theta_2^\star]\right\} \\ &= \exp\left\{-\frac{1}{4} h_\star^T [\Theta_1^\star + \Theta_2^\star] h_\star\right\}. \end{aligned} \quad (16)$$

Note that in the *symmetric* case (where  $\mathcal{U}_1 = \mathcal{U}_2$ ), there always exists a saddle point of  $\Psi$  with  $\Theta_1^\star = \Theta_2^\star$ , and the test  $\mathcal{T}_\star^1$  associated with such saddle point is quite transparent: it is the maximum likelihood test for two Gaussian distributions,  $\mathcal{N}(\theta_1^\star, \Theta_\star)$ ,  $\mathcal{N}(\theta_2^\star, \Theta_\star)$ , where  $\Theta_\star$  is the common value

of  $\Theta_1^*$  and  $\Theta_2^*$ , and the bound  $\epsilon_*$  for the risk of the test is nothing but the Hellinger affinity of these two Gaussian distributions, or, equivalently,

$$\epsilon_* = \exp \left\{ -\frac{1}{8} [\theta_1^* - \theta_2^*]^T \Theta_*^{-1} [\theta_1^* - \theta_2^*] \right\}. \quad (17)$$

Applying Proposition 3.2, we arrive at the following result:

**Proposition 3.3** *In the symmetric sub-Gaussian case (i.e., in the case of (12) with  $\mathcal{U}_1 = \mathcal{U}_2$ ), saddle point problem (13), (14) admits a saddle point of the form  $(h_*; \theta_1^*, \Theta_*, \theta_2^*, \Theta_*)$ , and the associated affine detector and its risk are given by*

$$\begin{aligned} \phi_*(\omega) &= h_*^T [\omega - w_*], \quad w_* = \frac{1}{2} [\theta_1^* + \theta_2^*]; \\ \epsilon_* &= \exp \left\{ -\frac{1}{8} [\theta_1^* - \theta_2^*]^T \Theta_*^{-1} [\theta_1^* - \theta_2^*] \right\}. \end{aligned} \quad (18)$$

As a result, when deciding, via  $\omega^K$ , on “sub-Gaussian hypotheses”  $H_S[\mathbf{R}^d, \mathcal{M}_\chi, \mathcal{M}_\chi]$ ,  $\chi = 1, 2$  (in fact - even on weaker hypotheses  $H_R[\mathbf{R}^d, \mathcal{M}_\chi, \mathcal{M}_\chi]$ ,  $\chi = 1, 2$ ), the risk of the test  $\mathcal{T}_*^K$  associated with  $\phi_*^{(K)}(\omega^K) := \sum_{t=1}^K \phi_*(\omega_t)$  is at most  $\epsilon_*^K$ .

In the symmetric single-observation Gaussian case, that is, when  $\mathcal{U}_1 = \mathcal{U}_2$  and we apply the test  $\mathcal{T}_* = \mathcal{T}_*^1$  to observation  $\omega \equiv \omega_1$  in order to decide on the hypotheses  $\mathcal{H}_\chi^G$ ,  $\chi = 1, 2$ , the above risk bound can be improved:

**Proposition 3.4** *Consider symmetric case  $\mathcal{U}_1 = \mathcal{U}_2 = \mathcal{U}$ , let  $(h_*; \theta_1^*; \Theta_1^*, \theta_2^*, \Theta_2^*)$  be “symmetric” - with  $\Theta_1^* = \Theta_2^* = \Theta_*$  - saddle point of function  $\Psi$  given by (13), and let  $\phi_*$  be the affine detector given by (15) and (16):*

$$\phi_*(\omega) = h_*^T [\omega - w_*], \quad h_* = \frac{1}{2} \Theta_*^{-1} [\theta_1^* - \theta_2^*], \quad w_* = \frac{1}{2} [\theta_1^* + \theta_2^*].$$

Let also

$$\delta = \sqrt{h_*^T \Theta_* h_*} = \frac{1}{2} \sqrt{[\theta_1^* - \theta_2^*]^T \Theta_*^{-1} [\theta_1^* - \theta_2^*]}, \quad (19)$$

so that

$$\delta^2 = h_*^T [\theta_1^* - w_*] = h_*^T [w_* - \theta_2^*] \text{ and } \epsilon_* = \exp \left\{ -\frac{1}{2} \delta^2 \right\}. \quad (20)$$

Let, further,  $\alpha \leq \delta^2$ ,  $\beta \leq \delta^2$ . Then

$$\begin{aligned} (a) \quad & \forall (\theta \in U_1, \Theta \in \mathcal{U}) : \text{Prob}_{\omega \sim \mathcal{N}(\theta, \Theta)} \{ \phi_*(\omega) \leq \alpha \} \leq \text{Erf}(\delta - \alpha/\delta) \\ (b) \quad & \forall (\theta \in U_2, \Theta \in \mathcal{U}) : \text{Prob}_{\omega \sim \mathcal{N}(\theta, \Theta)} \{ \phi_*(\omega) \geq -\beta \} \leq \text{Erf}(\delta - \beta/\delta), \end{aligned} \quad (21)$$

where

$$\text{Erf}(s) = \frac{1}{\sqrt{2\pi}} \int_s^\infty \exp\{-r^2/2\} dr$$

is the normal error function. In particular, when deciding, via a single observation  $\omega$ , on Gaussian hypotheses  $H_\chi^G$ ,  $\chi = 1, 2$ , with  $H_\chi^G$  stating that  $\omega \sim \mathcal{N}(\theta, \Theta)$  with  $(\theta, \Theta) \in U_\chi \times \mathcal{U}$ , the risk of the test  $\mathcal{T}_*^1$  associated with  $\phi_*$  is at most  $\text{Erf}(\delta)$ .

**Proof.** Let us prove (a) (the proof of (b) is completely similar). For  $\theta \in U_1$ ,  $\Theta \in \mathcal{U}$  we have

$$\begin{aligned} & \text{Prob}_{\omega \sim \mathcal{N}(\theta, \Theta)} \{ \phi_*(\omega) \leq \alpha \} = \text{Prob}_{\omega \sim \mathcal{N}(\theta, \Theta)} \{ h_*^T [\omega - w_*] \leq \alpha \} \\ &= \text{Prob}_{\xi \sim \mathcal{N}(0, I)} \{ h_*^T [\theta + \Theta^{1/2} \xi - w_*] \leq \alpha \} \\ &= \text{Prob}_{\xi \sim \mathcal{N}(0, I)} \{ [\Theta^{1/2} h_*]^T \xi \leq \alpha - \underbrace{h_*^T [\theta - w_*]}_{\substack{\geq h_*^T [\theta_1^* - w_*] = \delta^2 \\ \text{by (15.b), (20)}}} \} \leq \text{Prob}_{\xi \sim \mathcal{N}(0, I)} \{ [\Theta^{1/2} h_*]^T \xi \leq \alpha - \delta^2 \} \\ &= \text{Erf}([\delta^2 - \alpha]/\|\Theta^{1/2} h_*\|_2) \\ &\leq \text{Erf}([\delta^2 - \alpha]/\|\Theta_*^{1/2} h_*\|_2) \text{ [since } \delta^2 - \alpha \geq 0 \text{ and } h_*^T \Theta h_* \leq h_*^T \Theta_* h_* \text{ by (15.c)]} \\ &= \text{Erf}([\delta^2 - \alpha]/\delta). \end{aligned}$$

The “in particular” part of Proposition is readily given by (21) as applied with  $\alpha = \beta = 0$ .  $\square$

### 3.3 Testing multiple hypotheses from repeated observations

Consider the situation as follows: we are given

- observation space  $\Omega = \mathbf{R}^d$  and a symmetric w.r.t. the origin closed convex set  $\mathcal{H}$ ;
- $J$  closed convex sets  $\mathcal{M}_j \subset \mathbf{R}^{n_j}$ ,  $j = 1, \dots, J$ , along with  $J$  convex-concave continuous functions  $\Phi_j(h; \mu^j) : \mathcal{H} \times \mathcal{M}_j \rightarrow \mathbf{R}$ .

These data give rise to  $J$  hypotheses  $H_j = H_{\mathcal{R}}[\mathcal{H}, \mathcal{M}_j, \Phi_j]$  on the distribution of repeated observation  $\omega^\infty = \{\omega_t \in \mathbf{R}^d, t \geq 1\}$ . On the top of it, assume we are given a closeness relation – a subset  $\mathcal{C} \subset \{1, \dots, J\}^2$  which we assume to contain the diagonal  $((j, j) \in \mathcal{C} \text{ for all } j \leq J)$  and to be symmetric  $((i, j) \in \mathcal{C} \text{ if and only if } (j, i) \in \mathcal{C})$ . In the sequel, we interpret indexes  $i, j$  with  $(i, j) \in \mathcal{C}$  (and the hypotheses  $H_i, H_j$  with these indexes) as  $\mathcal{C}$ -close to each other.

Our goal is, given a positive integer  $K$  and  $K$ -repeated observation  $\omega^K = (\omega_1, \dots, \omega_K)$ , to decide, “up to closeness  $\mathcal{C}$ ,” on the hypotheses  $H_j$ ,  $j = 1, \dots, J$  (which is convenient to be thought of as hypotheses about the distribution of  $\omega^K$ ).

Let us act as follows<sup>3</sup>. Let us make

**Assumption II** For every pair  $i, j$ ,  $1 \leq i < j \leq J$ , with  $(i, j) \notin \mathcal{C}$ , the convex-concave function

$$\Psi_{ij}(h; \mu_i, \mu_j) = \frac{1}{2} [\Phi_i(-h; \mu_i) + \Phi_j(h; \mu_j)] : \mathcal{H} \times (\mathcal{M}_i \times \mathcal{M}_j) \rightarrow \mathbf{R}$$

has a saddle point  $(h^{ij}; \mu_i^{ij}, \mu_j^{ij})$  on  $\mathcal{H} \times (\mathcal{M}_i \times \mathcal{M}_j)$  (min in  $h$ , max in  $\mu_i, \mu_j$ ),

and let us set

$$\left. \begin{aligned} \epsilon_{ij} &= \exp\{\Psi_{ij}(h^{ij}; \mu_i^{ij}, \mu_j^{ij})\}, \\ \phi_{ij}(\omega) &= [h^{ij}]^T \omega + a_{ij} \\ a_{ij} &= \frac{1}{2} [\Phi_i(-h^{ij}; \mu_i^{ij}) - \Phi_j(h^{ij}; \mu_j^{ij})] \end{aligned} \right\}, 1 \leq i < j \leq J \text{ \& } (i, j) \notin \mathcal{C} \quad (22)$$

(cf. (8) and (9)). We further set

$$\phi_{ij}(\cdot) \equiv 0, \epsilon_{ij} = 1 \quad \forall (i, j) \in \mathcal{C},$$

and

$$\phi_{ij}(\omega) = -\phi_{ji}(\omega), \epsilon_{ij} = \epsilon_{ji}, 1 \leq j < i \leq J \text{ \& } (i, j) \notin \mathcal{C},$$

and set

$$\phi_{ij}^{(K)}(\omega^K) = \sum_{t=1}^K \phi_{ij}(\omega_t), 1 \leq i, j \leq J.$$

Now, by construction,

$$\phi_{ij}^{(K)}(\cdot) = -\phi_{ji}^{(K)}(\cdot), \epsilon_{ij} = \epsilon_{ji}, 1 \leq i, j \leq J.$$

Observe that for every  $j \leq J$  such that the  $K$ -repeated observation  $\omega^K$  obeys hypothesis  $H_j = H_{\mathcal{R}}[\mathcal{H}, \mathcal{M}_j, \Phi_j]$ , we have

$$\mathbf{E} \left\{ \exp\{\phi_{ij}^{(K)}(\omega^K)\} \right\} \leq \epsilon_{ij}^K, i = 1, \dots, J. \quad (23)$$

---

<sup>3</sup>The construction we are about to present and the first related result (Proposition 3.5) originate from [13, Section 3]; we reproduce them below to make our exposition self-contained.

(we have used (11) along with  $\phi_{ij} \equiv -\phi_{ji}$ ).

Furthermore, whenever  $[\alpha_{ij}]_{i,j \leq J}$  is a skew-symmetric matrix (i.e.,  $\alpha_{ij} = -\alpha_{ji}$ ), the shifted detectors

$$\widehat{\phi}_{ij}^{(K)}(\omega^K) = \phi_{ij}^{(K)}(\omega^K) + \alpha_{ij} \quad (24)$$

satisfy the scaled version of (23), specifically, for every  $j \leq J$  such that the  $K$ -repeated observation  $\omega^K$  obeys hypothesis  $H_j = H_{\mathcal{R}}[\mathbf{R}^d, \mathcal{M}_j, \Phi_j]$ , we have

$$\mathbf{E} \left\{ \exp\{\widehat{\phi}_{ij}^{(K)}(\omega^K)\} \right\} \leq \epsilon_{ij}^K \exp\{\alpha_{ij}\}, \quad i = 1, 2, \dots, J. \quad (25)$$

The bottom line is as follows:

**Proposition 3.5** *In the situation in question, given a closeness  $\mathcal{C}$ , consider the following test  $\mathcal{T}_{\mathcal{C}}^K$  deciding on the hypotheses  $H_j$ ,  $1 \leq j \leq J$ , on the distribution of  $K$ -repeated observation  $\omega^K$ : given skew-symmetric shift matrix  $[\alpha_{ij}]_{i,j}$  and observation  $\omega^K$ ,  $\mathcal{T}_{\mathcal{C}}^K$  accepts all hypotheses  $H_i$  such that*

$$\widehat{\phi}_{ij}^{(K)}(\omega^K) > 0 \text{ whenever } (i, j) \notin \mathcal{C} \quad (*_i) \quad (26)$$

and reject all hypotheses  $H_i$  for which the predicate  $(*_i)$  does not take place. Then

(i) Test  $\mathcal{T}_{\mathcal{C}}^K$  accepts some of (perhaps, none of) hypotheses  $H_i$ ,  $i = 1, \dots, J$ , and all accepted hypotheses, if any, are  $\mathcal{C}$ -close to each other. Besides,  $\mathcal{T}_{\mathcal{C}}^K$  has  $\mathcal{C}$ -risk at most

$$\widehat{\epsilon} = \max_i \sum_{j: (i,j) \notin \mathcal{C}} \epsilon_{ij}^K \exp\{-\alpha_{ij}\},$$

meaning that for every  $j_* \leq J$  such that the distribution  $\bar{P}_K$  of  $\omega^K$  obeys the hypothesis  $H_{j_*}$  (i.e., the hypothesis  $H_{j_*}$  is true), the  $\bar{P}_K$ -probability of the event

”either the true hypothesis  $H_{j_*}^K$  is not accepted, or the list of accepted hypotheses contains a hypothesis which is not  $\mathcal{C}$ -close to  $H_{j_*}$ ”

does not exceed  $\widehat{\epsilon}$ .

(ii) The infimum of  $\widehat{\epsilon}$  over all skew-symmetric shifts  $\alpha_{ij}$  is exactly the spectral norm  $\|E^{(K)}\|_{2,2}$  of the symmetric entry-wise nonnegative matrix

$$E^{(K)} = \left[ E_{ij}^{(K)} = \begin{cases} \epsilon_{ij}^K, & (i, j) \notin \mathcal{C} \\ 0, & (i, j) \in \mathcal{C} \end{cases} \right]_{i,j=1}^J. \quad (26)$$

This infimum is attained when the Perron-Frobenius eigenvector  $g$  of  $E^{(K)}$  can be selected to be positive, in which case an optimal selection of  $\alpha_{ij}$  is the selection

$$\alpha_{ij} = \ln(g_i/g_j), \quad 1 \leq i, j \leq J.$$

**Proof.** Given  $\omega^K$ , the test  $\mathcal{T}_{\mathcal{C}}^K$  can accept hypotheses  $H_i$  and  $H_j$  with  $(i, j) \notin \mathcal{C}$  only when  $\phi_{ij}^{(K)}(\omega^K) > 0$  and  $\phi_{ji}^{(K)}(\omega^K) > 0$ , which is impossible due to  $\phi_{ij}^{(K)}(\cdot) = -\phi_{ji}^{(K)}(\cdot)$ . Thus,  $\mathcal{T}_{\mathcal{C}}^K$  can accept  $H_i$  and  $H_j$  only when  $(i, j) \in \mathcal{C}$ . Further, let the distribution  $\bar{P}_K$  of  $\omega^K$  obey hypothesis  $H_{j_*}$ . Invoking (25) and the relation  $\widehat{\phi}_{j_*j}^{(K)}(\cdot) = -\widehat{\phi}_{jj_*}^{(K)}(\cdot)$ , for every  $j \leq J$  with  $(j_*, j) \notin \mathcal{C}$ , the  $\bar{P}_K$ -probability of the event  $\widehat{\phi}_{j_*j}^{(K)}(\omega^K) \leq 0$ , or, which is the same, of the event  $\widehat{\phi}_{jj_*}^{(K)}(\omega^K) \geq 0$ , is at most  $\epsilon_{jj_*}^K \exp\{\alpha_{jj_*}\} =$



$\epsilon_{j_*j}^K \exp\{-\alpha_{j_*j}\}$ . Using the union bound, the  $\bar{P}_K$ -probability of the event “ $H_{j_*}$  is not accepted” is at most

$$\sum_{j:(j_*j) \notin \mathcal{C}} \epsilon_{j_*j}^K \exp\{-\alpha_{j_*j}\} = \sum_j E_{j_*j}^{(K)} \exp\{-\alpha_{j_*j}\} \leq \hat{\epsilon}.$$

By construction of the test, when  $H_{j_*}$  is accepted and  $j$  is not  $\mathcal{C}$ -close to  $j_*$ ,  $H_j$  is not accepted (we have already seen that the test never accepts a pair of hypotheses which are not  $\mathcal{C}$ -close to each other). Thus, the  $\mathcal{C}$ -risk of  $\mathcal{T}_{\mathcal{C}}^K$  indeed is at most  $\hat{\epsilon}$ . Now,  $E^{(K)}$  is a symmetric entry-wise nonnegative matrix, so that its leading eigenvector  $g$  can be selected to be nonnegative. When  $g$  is positive, setting  $\alpha_{ij} = \ln(g_i/g_j)$ , we get for every  $i$

$$\sum_j E_{ij}^{(K)} \exp\{-\alpha_{ij}\} = \sum_j E_{ij}^{(K)} g_j/g_i = (E^{(K)}g)_i/g_i = \|E^{(K)}\|_{2,2},$$

and thus  $\hat{\epsilon} = \|E^{(K)}\|_{2,2}$ . The fact that this is the smallest possible, over skew-symmetric shifts  $\alpha_{ij}$ , value of  $\hat{\epsilon}$  is proved in [13]. When  $g$  is just nonnegative, consider close to  $E^{(K)}$  symmetric matrix  $\hat{E}$  with positive entries  $\hat{E}_{ij} \geq E_{ij}^{(K)}$ ; utilizing the (automatically strictly positive) Perron-Frobenius eigenvector  $g$  of this matrix, we, as was just explained, get skew-symmetric shifts  $\alpha_{ij}$  such that

$$\sum_j \hat{E}_{ij} \exp\{-\alpha_{ij}\} \leq \|\hat{E}\|_{2,2}$$

for all  $i$ ; the left hand side in this inequality is  $\geq \sum_j E_{ij}^{(K)} \exp\{-\alpha_{ij}\}$ , and the right hand side can be made arbitrarily close to  $\|E^{(K)}\|_{2,2}$  by making  $\hat{E}$  close enough to  $E^{(K)}$ . Thus, we indeed can make  $\hat{\epsilon}$  arbitrarily close to  $\|E^{(K)}\|_{2,2}$ .  $\square$

### 3.3.1 Special case: inferring colors

Assume that we are given  $J$  collections  $(\mathcal{H}, \mathcal{M}_j, \Phi_j)$ ,  $1 \leq j \leq J$ , of regular data with common observation space  $\Omega = \mathbf{R}^d$  and common  $\mathcal{H}$ , and thus have at our disposal  $J$  hypotheses  $H_j = H_{\mathcal{R}}[\mathcal{H}, \mathcal{M}_j, \Phi_j]$ , on the distribution of  $K$ -repeated observation  $\omega^K$ . Let also the index set  $\{1, \dots, J\}$  be partitioned into  $L \geq 2$  non-overlapping nonempty subsets  $\mathcal{I}_1, \dots, \mathcal{I}_L$ ; it is convenient to think about  $\ell$ ,  $1 \leq \ell \leq L$ , as the common color of indexes  $j \in \mathcal{I}_\ell$ , and that the colors of indexes  $j$  are inherited by the hypotheses  $H_j$ . The Color Inference (CI) problem we want to solve amounts to decide, given  $K$ -repeated observation  $\omega^K$  obeying one or more of the hypotheses  $H_1, \dots, H_J$ , on the color of these hypotheses. Note that it may happen that the distribution of  $\omega^K$  obeys a pair of hypotheses  $H_i, H_j$  of different colors. If it is not the case – that is, no distribution of  $\omega^K$  obeys a pair of hypotheses  $H_i, H_j$  of two distinct colors – we call the CI problem well-posed. In the well-posed case, we can speak about the color of the distribution of  $\omega^K$  provided this distribution obeys the union, over  $j = 1, \dots, J$ , of hypotheses  $H_j$ ; this is the color of (any of) the hypotheses  $H_j$  obeyed by the distribution of  $\omega^K$ , and the CI problem is to infer this color given  $\omega^K$ .

In order to process the CI problem via our machinery, let us define closeness  $\mathcal{C}$  as follows:

$$(i, j) \in \mathcal{C} \Leftrightarrow i \text{ and } j \text{ are of the same color.}$$

Assuming that the resulting  $\mathcal{C}$  ensures validity of Assumption II, we can apply the above scheme to build test  $\mathcal{T}_{\mathcal{C}}^K$ . We can then convert this test into a color inference as follows. Given a  $K$ -repeated observation  $\omega^K$ , it may happen that  $\mathcal{T}_{\mathcal{C}}^K$ , as applied to  $\omega^K$ , accepts one or more among the hypotheses  $H_j$ . In this case, by item (i) of Proposition 3.5, all accepted hypotheses are  $\mathcal{C}$ -close to each other (in other words, are of the same color), and we claim that this is the color of the distribution of  $K$ -repeated

observation we are dealing with. And if  $\mathcal{T}_C^K$ , as applied to  $\omega^K$ , accepts nothing, we claim that the color we are interested in remains undecided.

Let us analyze the just described color inferring procedure, let it be denoted  $\mathcal{A}^K$ . Observe, first, that in the situation in question, assuming w.l.o.g. that the sets  $\mathcal{I}_1, \mathcal{I}_2, \dots, \mathcal{I}_L$  are consecutive fragments in  $\{1, \dots, J\}$ , the matrix  $E^{(K)}$  given by (26) is naturally partitioned into  $L \times L$  blocks  $E^{pq} = (E^{qp})^T$ ,  $1 \leq p, q \leq L$ , where  $E^{pq}$  is comprised of entries  $E_{ij}^{(K)}$  with  $i \in \mathcal{I}_p, j \in \mathcal{I}_q$ . By construction of  $E^{(K)}$ , the diagonal blocks  $E^{pp}$  are zero, and off-diagonal blocks are entry-wise positive (since  $\epsilon_{ij}$  clearly is positive for all pairs  $i, j$  of different colors). It follows that Perron-Frobenius eigenvectors of  $E^{(K)}$  are strictly positive. This implies that for properly selected shifts  $\alpha_{ij} = -\alpha_{ji}$ , the quantity  $\hat{\epsilon}$  in Proposition 3.5 is equal to  $\|E^{(K)}\|_{2,2}$ ; in what follows we assume that the test  $\mathcal{T}_C^K$  utilizes exactly these optimal shifts, so that we are in the case of  $\hat{\epsilon} = \|E^{(K)}\|_{2,2}$ .

Now, it may happen (“bad case”) that that  $\|E^{(K)}\|_{2,2} \geq 1$ ; in this case Proposition 3.5 says nothing meaningful about the quality of the test  $\mathcal{T}_C^K$ , and consequently, we cannot say much about the quality of  $\mathcal{A}^K$ . In contrast to this, we claim that

**Lemma 3.1** *Assume that  $\hat{\epsilon} := \|E^{(K)}\|_{2,2} < 1$ . Then the CI problem is well posed, and whenever the distribution  $\bar{P}_K$  of  $\omega^K$  obeys one of the hypotheses  $H_j$ ,  $j = 1, \dots, J$ ,  $\mathcal{A}^K$  recovers correctly the color of  $\bar{P}_K$  with  $\bar{P}_K$ -probability at least  $1 - \hat{\epsilon}$ .*

The proof is immediate. In the good case, all entries in  $E^{(K)}$  are of magnitude  $< 1$ , whence  $\epsilon_{ij} < 1$  whenever  $(i, j) \notin \mathcal{C}$ , see (26), so that

$$\bar{\epsilon} := \max_{i,j} \{\epsilon_{ij} : (i, j) \notin \mathcal{C}\} < 1. \quad (27)$$

It follows that the nonzero entries in  $E^{(M)}$  are nonnegative and  $\leq \bar{\epsilon}^M$ , whence

$$\hat{\epsilon}(M) := \|E^{(M)}\|_{2,2} \leq J\bar{\epsilon}^M \rightarrow 0, \text{ as } M \rightarrow \infty. \quad (28)$$

In particular, for properly selected  $M$  we have  $\hat{\epsilon}(M) < 1/2$ . Applying Proposition 3.5 with  $M$  in the role of  $K$ , we see that if the distribution  $\bar{P}_K$  of  $\omega^K$  obeys hypothesis  $H_{j_*}$  with some  $j_* \leq J$  (due the origin of our hypotheses, this is exactly the same as to say that the distribution  $\bar{P}_M$  of  $\omega^M$  obeys  $H_{j_*}$ ), then with  $\bar{P}_M$ -probability at least  $1 - \hat{\epsilon}(M) > 1/2$  the test  $\mathcal{T}_C^M$  accepts hypothesis  $H_{j_*}$ . It follows that if  $\bar{P}_M$  obeys both  $H_{j'}$  and  $H_{j''}$ , then  $\mathcal{T}_C^M$  will accept  $H_{j'}$  and  $H_{j''}$  simultaneously with positive  $\bar{P}_M$ -probability, and since, as we have already explained,  $\mathcal{T}_C^M$  never accepts two hypotheses of different color simultaneously, we conclude that  $H_{j'}$  and  $H_{j''}$  are of the same color. This conclusion holds true whenever the distribution of  $\omega^K$  obeys one or more of the hypotheses  $H_j$ ,  $1 \leq j \leq K$ , meaning that the CI problem is well posed.

Invoking Proposition 3.5, we conclude that if the distribution  $\bar{P}_K$  of  $\omega^K$  obeys, for some  $j_*$ , the hypothesis  $H_{j_*}$ , then the  $\bar{P}_K$ -probability for  $\mathcal{T}_C^K$  to accept  $H_{j_*}$  is at least  $1 - \hat{\epsilon}(K)$ ; and from the preceding analysis, whenever  $\mathcal{T}_C^K$  accepts  $H_{j_*}$  such that  $\bar{P}_K$  obeys  $H_{j_*}$ ,  $\mathcal{A}^K$  correctly infers the color of  $\bar{P}_K$ , as claimed.  $\square$

Finally, we remark that when (27) holds, (28) implies that  $\hat{\epsilon}(K) \rightarrow 0$  as  $K \rightarrow \infty$ , so that the CI problem is well posed, and for every desired risk level  $\epsilon \in (0, 1)$  we can find efficiently observation time  $K = K(\epsilon)$  such that  $\hat{\epsilon}(K) \leq \epsilon$ . As a result, for this  $K$  the color inferring procedure  $\mathcal{A}_K$  recovers the color of the distribution  $\bar{P}_K$  of  $\omega^K$  (provided this distribution obeys some of the hypotheses  $H_1, \dots, H_J$ ) with  $\bar{P}_K$ -probability at least  $1 - \epsilon$ .

## 4 Application: aggregating estimates by testing

Let us consider the situation as follows:

- We are given  $I$  triples of regular data  $(\mathcal{H}, \mathcal{M}_i, \Phi_i)$ ,  $1 \leq i \leq I$ , with common  $\mathcal{H}$  and  $\Omega = \mathbf{R}^d$  and the parameter sets  $\mathcal{M}_i$  sharing the common embedding space  $\mathbf{R}^n$ ; for the sake of simplicity, assume that  $\mathcal{M}_i$  are bounded (and thus are nonempty convex compact sets in  $\mathbf{R}^n$ ) and the continuous convex-concave functions  $\Phi_i(h; \mu) : \mathcal{H} \times \mathcal{M}_j \rightarrow \mathbf{R}$  are coercive in  $H$ :  $\Phi_i(h_t, \mu) \rightarrow \infty$  whenever  $\mu \in \mathcal{M}_i$  and sequence  $\{h_t \in \mathcal{H}, t \geq 1\}$  satisfies  $\|h_t\|_2 \rightarrow \infty, t \rightarrow \infty$ .
- We observe a realization of  $K$ -repeated observation  $\omega^K = (\omega_1, \dots, \omega_K)$  with i.i.d.  $\omega_t$ 's drawn from unknown probability distribution  $\bar{P}$  known to belong to the family  $\mathcal{P} = \bigcup_{i \leq I} \mathcal{S}[\mathcal{H}, \mathcal{M}_i, \Phi_i]$ .

Thus, “in the nature” there exists  $\bar{i} \leq I$  and  $\bar{\mu} \in \mathcal{M}_{\bar{i}}$  such that

$$\ln(\mathbf{E}_{\omega \sim \bar{P}}\{\exp\{h^T \omega\}\}) \leq \Phi_{\bar{i}}(h, \bar{\mu}) \quad \forall h \in \mathcal{H}. \quad (29)$$

we call  $\bar{\mu}$  the parameter associated with  $\bar{P}$ , and our goal is to recover from  $K$ -repeated observation  $\omega^K$  the image  $\bar{g} = G\bar{\mu}$  of  $\bar{\mu}$  under a given linear mapping  $\mu \mapsto G\mu : \mathbf{R}^n \rightarrow \mathbf{R}^m$ .

Undoubtedly, parameter estimation problem is a fundamental problem of mathematical statistics, and as such is the subject of a huge literature. In particular, several constructions of estimators based on testing of convex hypotheses have been studied in connection with signal reconstruction [4, 5, 6] and linear functionals estimation [10, 11]. Our actual goal to be addressed below is more modest: we assume that we are given  $L$  candidate estimates  $g_1, \dots, g_L$  of  $\bar{g}$  (these estimates could be outputs of various estimation routines applied to independent observations sampled from  $\bar{P}$ ), and our goal is to select the best – the  $\|\cdot\|_2$ -closest to  $\bar{g}$  among the estimates  $g_1, \dots, g_L$ . This is the well-known problem of *aggregating estimates*, and our goal is to process this aggregation problem via the Color Inference procedure from section 3.3.1.

#### 4.1 Aggregation procedure

It should be stressed that as stated, the aggregation problem appears to be ill-posed: there could be several pairs  $(\bar{i}, \bar{\mu})$  satisfying (29), and the values of  $G\mu$  at the  $\mu$ -components of these pairs could be different for different pairs, so that  $\bar{g}$  not necessary is well defined. One way to resolve this ambiguity would be to assume that given  $\bar{P} \in \mathcal{P}$ , relation (29) uniquely defines  $\bar{\mu}$ . We, however, prefer another setting:  $\bar{\mu}$  and  $\bar{i}$  satisfying (29), same as  $\bar{P} \in \mathcal{P}$ , are “selected by nature” (perhaps, from several alternatives), and the performance of the aggregating procedure we are about to develop should be independent of what is the nature’s selection.

When processing the aggregation problem, we assume w.l.o.g. that all points  $g_1, \dots, g_L$  are distinct from each other. Let us split the space  $\mathbf{R}^m$  where  $G\mu$  takes values into  $L$  Voronoi cells

$$\begin{aligned} V_\ell &= \{g \in \mathbf{R}^m : \|g - g_\ell\|_2 \leq \|g - g_{\ell'}\|_2 \quad \forall \ell' \leq L\} \\ &= \{g \in \mathbf{R}^m : u_{\ell\ell'}^T g \leq v_{\ell\ell'} \quad \forall (\ell' \leq L, \ell' \neq \ell)\}, \\ u_{\ell\ell'} &= \|g_{\ell'} - g_\ell\|_2^{-1} [g_{\ell'} - g_\ell], \quad v_{\ell\ell'} = \frac{1}{2} u_{\ell\ell'}^T [g_{\ell'} + g_\ell], \quad 1 \leq \ell, \ell' \leq L, \ell \neq \ell'. \end{aligned} \quad (30)$$

Note that  $V_\ell$  is comprised of all points  $g$  from  $\mathbf{R}^m$  for which  $g_\ell$  is (one of) the  $\|\cdot\|_2$ -closest to  $g$  among the points  $g_1, \dots, g_L$ . Let us set

$$W_\ell^i = \{\mu \in \mathcal{M}_i : G\mu \in V_\ell\}, \quad 1 \leq i \leq I, 1 \leq \ell \leq L, \quad (31)$$

so that  $W_\ell^i$  are convex compact sets in  $\mathbf{R}^n$ . Observe that  $g_\ell$  can be a solution to the aggregation problem (that is, the closest to  $\bar{g}$  point among  $g_1, \dots, g_L$ ) only when  $\bar{g} = G\bar{\mu}$  belongs to  $V_\ell$ , that is, only when  $\bar{\mu} \in W_\ell^{\bar{i}}$  for some  $\bar{i}$ , implying that at least one of the sets  $W_\ell^1, W_\ell^2, \dots, W_\ell^I$  is nonempty. Whether the latter condition is indeed satisfied for a given  $\ell$  this can be found out efficiently via solving  $I$  convex

feasibility problems. If the latter condition does *not* hold for some  $\ell$  (let us call the associated estimate  $g_\ell$  *redundant*), we can eliminate  $g_\ell$  from the list of estimates to be aggregated without affecting the solution to the aggregation problem. Then we can redefine the Voronoi cells for the reduced list of estimates in the role of our original list, check whether this list still contains a redundant estimate, eliminate the latter, if it exists, and proceed recursively until a list of estimates (which by construction still contains all solutions to the aggregation problem) with no redundant estimates is built. We lose nothing when assuming that this “purification” was carried out in advance, so that already the initial list  $g_1, \dots, g_L$  of estimates does not contain redundant ones. Of course, we lose nothing when assuming that  $L \geq 2$ . Thus, from now on we assume that  $L \geq 2$  and for every  $\ell \leq L$ , at least one of the sets  $W_\ell^i$ ,  $1 \leq i \leq I$ , is nonempty.

Note that to solve the aggregation problem to optimality is exactly the same, in terms of the sets  $W_\ell^i$ , as to identify  $\ell$  such that  $\bar{\mu} \in W_\ell^i$  for some  $i$ . We intend to reduce this task to solving a series of Color Inference problems. We start with presenting the principal building block of our construction – *Individual Inference procedure*.

**Individual Inference procedure** is parameterized by  $\ell \in \{1, \dots, L\}$  and a real  $\delta > 0$ . Given  $\ell$  and  $\delta$ , we initialize the algorithm as follows:

- mark as red all nonempty sets  $W_\ell^i$  along with their elements and the corresponding regular data  $(\mathcal{H}, W_\ell^i, \Phi_i|_{\mathcal{H} \times W_\ell^i})$ ;
- look one by one at all sets  $W_{\ell'}^i$  with  $i \leq I$  and  $\ell' \neq \ell$ , and associate with these sets their chunks

$$W_{\ell\ell'}^{i\delta} = \{\mu \in W_{\ell'}^i : u_{\ell\ell'}^T[G\mu] \geq v_{\ell\ell'} + \delta\}. \quad (32)$$

Note that the resulting sets are convex and compact. Whenever  $W_{\ell\ell'}^{i\delta}$  is nonempty, we mark blue this set along with its elements and the corresponding regular data  $(\mathcal{H}, W_{\ell\ell'}^{i\delta}, \Phi_i|_{\mathcal{H} \times W_{\ell\ell'}^{i\delta}})$ .

As a result of the above actions, we get a collection of nonempty convex compact subsets  $W_\ell^{s\delta}$ ,  $s = 1, \dots, S_\ell^\delta$ , of  $\mathbf{R}^d$  and associated regular data  $D_\ell^{s\delta} = (\mathcal{H}, W_\ell^{s\delta}, \Phi_\ell^{s\delta})$ ; the sets  $W_\ell^{s\delta}$ , same as their elements and regular data  $\mathcal{D}_\ell^{s\delta}$ , are colored in red and blue. Note that the sets  $W_\ell^{s\delta}$  of different colors do not intersect (since their images under the mapping  $G$  do not intersect), so that a point  $\mu \in \mathbf{R}^m$  gets at most one color. Note also that our collection definitely contains red components.

Individual Inference Procedure  $\mathcal{A}_K^{\ell\delta}$  infers the color of a regular data  $\mathcal{D} \in \mathbf{D}_\ell^\delta = \{\mathcal{D}_\ell^{s\delta}, 1 \leq s \leq S_\ell^\delta\}$  given i.i.d.  $K$ -repeated observation  $\omega^K$  drawn from a distribution  $P \in \mathcal{S}[\mathcal{D}]$ : when the collection  $\mathbf{D}_\ell^\delta$  contains both red and blue regular data,  $\mathcal{A}_K^{\ell\delta}$  is exactly Color Inference procedure from section 3.3.1 associated with this collection and our coloring<sup>4</sup>; if no blue regular data is present,  $\mathcal{A}_K^{\ell\delta}$  always infers that the color is red.

Observe that if the collection  $\mathbf{D}_\ell^\delta$  of regular data we have built contains no blue data for some value  $\bar{\delta}$  of  $\delta$ , the same holds true for all  $\delta \geq \bar{\delta}$ . Let us define the risk  $\hat{e}(\ell, \delta)$  of Individual Inference Procedure with parameters  $\ell, \delta$  as follows: when  $\delta$  is such that  $\mathbf{D}_\ell^\delta$  contains no blue regular data,  $\hat{e}(\ell, \delta) = 0$ , otherwise  $\hat{e}(\ell, \delta)$  is as stated in Proposition 3.5. Note that whenever  $\delta > 0$ ,  $\ell \leq L$ ,  $s \leq S_\ell^\delta$ ,  $\mu \in W_\ell^{s\delta}$  and a probability distribution  $P$  satisfies

$$\mathbf{E}_{\omega \sim P} \{\exp\{h^T \omega\}\} \leq \Phi^s(h, \mu) \quad \forall h \in \mathcal{H},$$

---

<sup>4</sup>The procedure is well defined, since by our assumption, all convex-concave functions we need to deal with are continuous, coercive in the minimization variable, and have closed convex minimization and compact convex maximization domains, so that the required saddle points do exist.

the quantity  $\widehat{\epsilon}(\ell, \delta)$ , by construction, is an upper bound on  $P$ -probability of the event “as applied to observation  $\omega^K = (\omega_1, \dots, \omega_K)$  with  $\omega_1, \dots, \omega_K$  drawn, independently of each other, from  $P$ ,  $\mathcal{A}_K^{\ell\delta}$  does not recover correctly the color of  $\mu$ .” Observe that  $\widehat{\epsilon}(\ell, \delta) = 0$  for large enough values of  $\delta$  (since for large  $\delta$  the collection  $\mathbf{D}_\ell^\delta$  contains no blue data; recall that the parameter sets  $\mathcal{M}_i$  are bounded). Besides this, we claim that  $\widehat{\epsilon}(\ell, \delta)$  is nonincreasing in  $\delta > 0$ .

To support the claim, assume that  $\delta' \geq \delta'' > 0$  are such that  $\mathbf{D}_{\ell'}^{\delta''}$  contains blue data, and let us show that  $\epsilon' := \widehat{\epsilon}(\ell, \delta') \leq \epsilon'' := \widehat{\epsilon}(\ell, \delta'')$ . Indeed, recall that  $\epsilon'$  and  $\epsilon''$  are  $\|\cdot\|_{2,2}$ -norms of respective symmetric entry-wise nonnegative matrices  $E', E''$  (see Proposition 3.5). When increasing  $\delta$  from  $\delta = \delta''$  to  $\delta = \delta'$ , we reduce the associated  $E$ -matrix to its submatrix<sup>5</sup> and further reduce the entries in this submatrix<sup>6</sup>, and thus reduce the norm of the matrix.

**Aggregation procedure** we propose is as follows:

Given tolerance  $\epsilon \in (0, \frac{1}{2})$ , for every  $\ell = 1, \dots, L$  we specify  $\delta_\ell > 0$ , the smaller the better, in such a way that  $\widehat{\epsilon}(\ell, \delta_\ell) \leq \epsilon/L$ <sup>7</sup>. Given observation  $\omega^K$ , we run the procedures  $\mathcal{A}_K^{\ell\delta_\ell}$ ,  $1 \leq \ell \leq L$ . Whenever  $\mathcal{A}_K^{\ell\delta_\ell}$  returns a color, we assign it to the index  $\ell$  and to the vector  $g_\ell$ , so that after all  $\ell$ 's are processed, some  $g_\ell$ 's get color “red,” some get color “blue,” and some do not get color. The aggregation procedure returns, as a solution  $\hat{g}(\omega^K)$ , a (whatever) red vector if one was discovered, and returns, say,  $g_1$  otherwise.

**Proposition 4.1** *In the situation and under the assumptions described in the beginning of section 4, let  $\omega^K = (\omega_1, \dots, \omega_K)$  be  $K$ -element i.i.d. sample drawn from a probability distribution  $\bar{P}$  which, taken along with some  $\bar{i} \leq I$  and  $\bar{\mu} \in \mathcal{M}_{\bar{i}}$ , satisfies (29). Then the  $\bar{P}$ -probability of the event*

$$\|G\bar{\mu} - \hat{g}(\omega^K)\|_2 \leq \min_{\ell \leq L} \|G\bar{\mu} - g_\ell\|_2 + 2 \max_{\ell} \delta_\ell \quad (33)$$

*is at least  $1 - \epsilon$ .*

**Simple illustration.** Let  $I = 1$ ,  $\mathcal{H} = \Omega = \mathbf{R}^d$ , and let  $\mathcal{M} = \mathcal{M}_1$  be a nonempty convex compact subset of  $\mathbf{R}^d$ . Further, suppose that  $\Phi(h; \mu) := \Phi_1(h; \mu) = h^T \mu + \frac{1}{2} h^T \Theta h : \mathcal{H} \times \mathcal{M} \rightarrow \mathbf{R}$ , where  $\Theta$  is a given positive definite matrix. We are also given a  $K$ -element i.i.d. sample  $\omega^K$  drawn from a sub-Gaussian distribution  $P$  with sub-Gaussianity parameters  $(\mu, \Theta)$ . Let also  $G\mu \equiv \mu$ , so that the aggregation problem we are interested in reads: “given  $L$  estimates  $g_1, \dots, g_L$  of the expectation  $\mu$  of a sub-Gaussian random vector  $\omega$  with sub-Gaussianity parameters  $(\mu, \Theta)$ , with a known  $\Theta \succ 0$ , and  $K$ -repeated i.i.d. sample  $\omega^K$  from the distribution of the vector, we want to select  $g_\ell$  which is  $\|\cdot\|_2$ -closest to the true expectation  $\mu$  of  $\omega$ .” From now on we assume that  $g_1, \dots, g_L \in \mathcal{M}$  (otherwise, projecting the estimates onto  $\mathcal{M}$ , we could provably improve their quality) and that  $g_1, \dots, g_L$  are distinct from each other.

<sup>5</sup>The sets  $W_{\ell\ell'}^{\delta}$  shrink as  $\delta$  grows, thus some of the blue sets  $W_{\ell\ell'}^{\delta}$  which are nonempty at  $\delta = \delta''$  can become empty when  $\delta$  increases to  $\delta'$ .

<sup>6</sup>Indeed, by Proposition 3.5, these entries are obtained from saddle point values of some convex-concave functions by a monotone transformation; it is immediately seen that as  $\delta$  grows, these functions and the domains of the minimization argument remain intact, while the domains of the maximization argument shrink, so that the saddle point values cannot increase.

<sup>7</sup>For instance, we could start with  $\delta = \delta^0$  large enough to ensure that  $\widehat{\epsilon}(\ell, \delta^0) = 0$ , and select  $\delta_\ell$  as either the last term in the progression  $\delta^i = \kappa^i \delta_0$ ,  $i = 0, 1, \dots$ , for some  $\kappa \in (0, 1)$ , such that  $\widehat{\epsilon}(\ell, \delta^i) \leq \epsilon/L$ , or the first term in this progression which is “negligibly small” (say, less than  $10^{-6}$ ), depending on what happens first.

In our situation, the sets (30), (31), (32) and functions  $\Phi^i$  become

$$\begin{aligned} W_\ell &= \{\mu \in \mathcal{M} : u_{\ell\ell'}^T \mu \leq v_{\ell\ell'}, \forall (\ell' \leq L : \ell' \neq \ell)\}, \\ u_{\ell\ell'} &= \frac{g_{\ell'} - g_\ell}{\|g_{\ell'} - g_\ell\|_2}, \quad v_{\ell\ell'} = \frac{1}{2}[g_\ell + g_{\ell'}], \ell \neq \ell'; \\ W_{\ell\ell'}^\delta &= \{\mu \in \mathcal{M} : u_{\ell\ell'}^T \mu \geq v_{\ell\ell'} + \delta\}, \quad 1 \leq \ell, \ell' \leq L, \ell \neq \ell', \\ \Phi(h; \mu) &= h^T \mu + \frac{1}{2} h^T \Theta h, \end{aligned}$$

(we are in the case of  $I = 1$  and thus suppress index  $i$  in the notation for  $W$ 's and  $\Phi$ ). Note that Individual Inference Procedure  $\mathcal{A}_K^{\ell\delta}$  deals with exactly one red hypothesis,  $H_{Si}[\mathbf{R}^d, W_\ell, \Phi]$ , and at most  $L - 1$  blue hypotheses  $H_{Si}[\mathbf{R}^d, W_{\ell\ell'}^\delta, \Phi]$  associated with nonempty sets  $W_{\ell\ell'}^\delta$  and  $\ell' \neq \ell$ .

Applying the construction from section 3.3.1, we arrive at the aggregation routine as follows (below  $\ell, \ell'$  vary in  $\{1, \dots, L\}$ ):

- We set

$$\begin{aligned} \delta_\ell &= \max_{\ell' \neq \ell} \sqrt{\ln \left( \frac{L\sqrt{L-1}}{\epsilon K} \right)} u_{\ell\ell'}^T \Theta u_{\ell\ell'}; \\ w_{\ell\ell'} &= \frac{1}{2} [g_\ell + g_{\ell'} + \delta_\ell u_{\ell\ell'}], \quad \ell \neq \ell'; \\ \psi_{\ell\ell'}^{(K)}(\omega^K) &= \frac{\delta_\ell}{2u_{\ell\ell'}^T \Theta u_{\ell\ell'}} u_{\ell\ell'}^T \left[ K w_{\ell\ell'} - \sum_{t=1}^K \omega_t \right] + \frac{1}{2} \ln(L-1), \quad \ell \neq \ell'. \end{aligned} \tag{34}$$

- Given  $\omega^K$ , for  $1 \leq \ell \leq L$  we assign vector  $g_\ell$  color “red,” if  $\psi_{\ell\ell'}^{(K)}(\omega^K) > 0$  for all  $\ell' \neq \ell$ , otherwise we do not assign  $g_\ell$  any color;
- If red vectors were found, we output (any) one of them as solution to the aggregation problem; if no red vectors are found, we output, say,  $g_1$  as solution.

Proposition 4.1 as applied to the situation in question states that whenever  $\omega_1, \omega_2, \dots, \omega_K$  are drawn, independently from each other, from a sub-Gaussian distribution  $P$  with parameters  $\mu, \Theta$ , then with  $P$ -probability at least  $1 - \epsilon$  the result  $\hat{g}(\omega^K)$  of the above aggregation routine satisfies the relation

$$\|\mu - \hat{g}(\omega^K)\|_2 \leq \min_{1 \leq \ell \leq L} \|\mu - g_\ell\|_2 + 2 \max_\ell \delta_\ell,$$

which essentially recovers the classical  $\ell_2$  oracle inequality (cf. [12, Theorem 4]).

## 5 Beyond the scope of affine detectors

### 5.1 Lifted detectors

The tests developed in sections 3.2.2 and 3.3 were based on *affine detectors* – affine functions  $\phi(\omega)$  associated with pairs of composite hypotheses  $H_1 : P \in \mathcal{P}_1, H_2 : P \in \mathcal{P}_2$  on the probability distribution  $P$  of observation  $\omega \in \Omega = \mathbf{R}^d$ . Such detectors were built to satisfy the relations

$$\int_\Omega \exp\{-\phi_*(\omega)\} P(d\omega) \leq \epsilon_* \quad \forall P \in \mathcal{P}_1 \quad \& \quad \int_\Omega \exp\{\phi_*(\omega)\} P(d\omega) \leq \epsilon_* \quad \forall P \in \mathcal{P}_2, \tag{35}$$

with as small  $\epsilon_*$  as possible (cf. (10)), and affinity of  $\phi$  is of absolutely no importance here: all constructions in sections 3.2.2, 3.3 were based upon availability of pairwise detectors  $\phi$ , *affine or not*, satisfying, for the respective pairs of composite hypotheses, relations (10) with some known  $\epsilon_*$ . So far, affinity of detectors was utilized only when *building* detectors satisfying (35) via the generic scheme presented in section 3.1.



Now note that given a random observation  $\zeta$  taking values in some  $\mathbf{R}^d$  along with a deterministic function  $Z(\zeta) : \mathbf{R}^d \rightarrow \mathbf{R}^D$ , we can convert an observation  $\zeta$  into an observation

$$\omega = (\zeta, Z(\zeta)) \in \mathbf{R}^d \times \mathbf{R}^D.$$

Here  $\omega$  is a deterministic transformation of  $\zeta$  which “remembers”  $\zeta$ , so that to make statistical inferences from observations  $\zeta$  is exactly the same as to make them from observations  $\omega$ . However, detectors which are affine in  $\omega$  can be nonlinear in  $\zeta$ : for instance, for  $Z(\zeta) = \zeta\zeta^T$ , affine in  $\omega$  detectors are exactly detectors *quadratic* in  $\zeta$ . We see that within the framework of our approach, passing from  $\zeta$  to  $\omega$  allows to consider a wider family of detectors and thus to arrive at a wider family of tests. The potential bottleneck here is the necessity to bring the “augmented” observations  $(\zeta, Z(\zeta))$  into the scope of our setup.

**Example: distributions with bounded support.** Consider the case where the distribution  $P$  of our observation  $\zeta \in \mathbf{R}^d$  belongs to a family  $\mathcal{P}$  of Borel probability distributions supported on a given bounded set, for the sake of simplicity – on the unit Euclidean ball  $B$  of  $\mathbf{R}^d$ . Given a continuous function  $Z(\cdot) : \mathbf{R}^d \rightarrow \mathbf{R}^D$ , our goal is to cover the family  $\mathcal{P}^+$  of distributions  $P^+[P]$  of  $\omega = (\zeta, Z(\zeta))$  induced by distributions  $P \in \mathcal{P}$  of  $\zeta$  by a family  $\mathcal{P}[\mathcal{H}, \mathcal{M}, \Phi]$  (or  $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$ ) associated with some regular data, thus making the machinery we have developed so far applicable to the family of distribution  $\mathcal{P}^+$ . Assuming w.l.o.g. that

$$\|Z(z)\|_2 \leq 1 \quad \forall (z : \|z\|_2 \leq 1),$$

observe that for  $P \in \mathcal{P}$  the distribution  $P^+[P]$  is sub-Gaussian with sub-Gaussianity parameters  $(\theta_P, \Theta = 2I_{d+D})$ , where

$$\theta_P = \int_B (\zeta, Z(\zeta)) P(d\zeta).$$

It follows that

*If we can point out a convex compact set  $U$  in  $\mathbf{R}^d \times \mathbf{R}^D$  such that*

$$\theta_P \in U \quad \forall P \in \mathcal{P}, \tag{36}$$

*then, specifying regular data as*

$$\begin{aligned} \mathcal{H} &= \Omega := \mathbf{R}^d \times \mathbf{R}^D, \\ \mathcal{M} &= U, \\ \Phi(h; \mu) &= h^T \mu + h^T h : \mathcal{H} \times \mathcal{M} \rightarrow \mathbf{R}, \end{aligned}$$

*we ensure that the family  $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$  contains the family of probability distributions  $\mathcal{P}^+ = \{P^+[P] : P \in \mathcal{P}\}$  on  $\Omega$ .*

How useful is this (by itself pretty crude) observation in the context of our approach depends on how much information on  $\mathcal{P}$  can be “captured” by a properly selected convex compact set  $U$  satisfying (36). We are about to consider in more details “quadratic lifting” – the case where  $Z(\zeta) = \zeta\zeta^T$ .

### 5.1.1 Quadratic lifting, Gaussian case

Consider the situation where we are given

- a nonempty bounded set  $U$  in  $\mathbf{R}^m$ ;



- a nonempty convex compact subset  $\mathcal{U}$  of the positive semidefinite cone  $\mathbf{S}_+^d$ ;
- a matrix  $\Theta_* \succ 0$  such that  $\Theta_* \succeq \Theta$  for all  $\Theta \in \mathcal{U}$ ;
- an affine mapping  $u \mapsto \mathcal{A}(u) = A[u; 1] : \mathbf{R}^m \rightarrow \Omega = \mathbf{R}^d$ , where  $A$  is a given  $d \times (m+1)$  matrix.

Now, a pair  $(u \in U, \Theta \in \mathcal{U})$  specifies Gaussian random vector  $\zeta \sim \mathcal{N}(\mathcal{A}(u), \Theta)$  and thus specifies a Borel probability distribution  $P[u, \Theta]$  of  $(\zeta, \zeta\zeta^T)$ . Let  $\mathcal{Q}(U, \mathcal{U})$  be the family of probability distributions on  $\Omega = \mathbf{R}^d \times \mathbf{S}^d$  stemming in this fashion from Gaussian distributions with parameters from  $U \times \mathcal{U}$ . Our goal is to cover the family  $\mathcal{Q}(U, \mathcal{U})$  by a family of the type  $\mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]$ , which, as it was already explained, would allow to use the machinery developed so far in order to decide on pairs of composite Gaussian hypotheses

$$\begin{aligned} H_1 : \zeta &\sim \mathcal{N}(u, \Theta) \text{ with } (u, \Theta) \in U_1 \times \mathcal{U}_1 \\ H_2 : \zeta &\sim \mathcal{N}(u, \Theta) \text{ with } (u, \Theta) \in U_2 \times \mathcal{U}_2 \end{aligned}$$

via tests based on detectors which are *quadratic* in  $\zeta$ .

It is convenient to represent a linear form on  $\Omega = \mathbf{R}^d \times \mathbf{S}^d$  as

$$h^T z + \frac{1}{2} \text{Tr}(HZ),$$

where  $(h, H) \in \mathbf{R}^d \times \mathbf{S}^d$  is the “vector of coefficients” of the form, and  $(z, Z) \in \mathbf{R}^d \times \mathbf{S}^d$  is the argument of the form.

We denote by  $b = [0; 0; \dots; 0; 1] \in \mathbf{R}^{m+1}$  the last basic orth of  $\mathbf{R}^{m+1}$ . We assume that for some  $\delta \geq 0$  it holds

$$\|\Theta^{1/2} \Theta_*^{-1/2} - I\| \leq \delta \quad \forall \Theta \in \mathcal{U}, \quad (37)$$

where  $\|\cdot\|$  is the spectral norm. Observe that for every  $\Theta \in \mathcal{U}$  we have  $0 \preceq \Theta_*^{-1/2} \Theta \Theta_*^{-1/2} \preceq I$ , whence  $\|\Theta^{1/2} \Theta_*^{-1/2}\| \leq 1$ , that is, (37) is always satisfied with  $\delta = 2$ . Thus, we can assume w.l.o.g. that  $\delta \in [0, 2]$ . Finally, we set  $B = \begin{bmatrix} A \\ b^T \end{bmatrix}$ .

A desired “covering” of  $\mathcal{Q}(U, \mathcal{U})$  is given by the following

**Proposition 5.1** *In the notation and under the assumptions of this section, let  $\gamma \in (0, 1)$  and a convex compact computationally tractable set  $\mathcal{Z} \subset \{W \in \mathbf{S}_+^{m+1} : W_{m+1, m+1} = 1\}$  be given, and let  $\mathcal{Z}$  contain all matrices  $Z(u) := [u; 1][u; 1]^T$  with  $u \in U$ . Denoting by  $\phi_{\mathcal{Z}}(\cdot)$  the support function of  $\mathcal{Z}$ :*

$$\phi_{\mathcal{Z}}(W) = \max_{Z \in \mathcal{Z}} \text{Tr}(ZW) : \mathbf{S}^{m+1} \rightarrow \mathbf{R},$$

let us set

$$\begin{aligned} \mathcal{H} &= \mathcal{H}_\gamma := \{(h, H) \in \mathbf{R}^d \times \mathbf{S}^d : -\gamma \Theta_*^{-1} \preceq H \preceq \gamma \Theta_*^{-1}\} \\ \mathcal{M} &= \mathcal{U}, \\ \Phi(h, H; \Theta) &= -\frac{1}{2} \ln \text{Det}(I - \Theta_*^{1/2} H \Theta_*^{1/2}) + \frac{1}{2} \text{Tr}([\Theta - \Theta_*]H) + \frac{\delta(2-\delta)}{2(1-\gamma)} \|\Theta_*^{1/2} H \Theta_*^{1/2}\|_F^2 \\ &\quad + \Gamma_{\mathcal{Z}}(h, H) : \mathcal{H} \times \mathcal{U} \rightarrow \mathbf{R}, \quad [\|\cdot\|_F \text{ is the Frobenius norm}] \\ \Gamma_{\mathcal{Z}}(h, H) &= \frac{1}{2} \phi_{\mathcal{Z}}(bh^T A + A^T h b^T + A^T H A + B^T [H, h]^T [\Theta_*^{-1} - H]^{-1} [H, h] B) \\ &= \frac{1}{2} \phi_{\mathcal{Z}} \left( B^T \left[ \begin{bmatrix} H & h \\ h^T & \end{bmatrix} + [H, h]^T [\Theta_*^{-1} - H]^{-1} [H, h] \right] B \right). \end{aligned} \quad (38)$$

Then  $\mathcal{H}, \mathcal{M}, \Phi$  form a regular data, and

$$\mathcal{Q}(U, \mathcal{U}) \subset \mathcal{S}[\mathcal{H}, \mathcal{M}, \Phi]. \quad (39)$$

Besides this, function  $\Phi(h, H; \Theta)$  is coercive in the convex argument: whenever  $(h_i, H_i) \in \mathcal{H}$  and  $\|(h_i, H_i)\| \rightarrow \infty$  as  $i \rightarrow \infty$ , we have  $\Phi(h_i, H_i; \Theta) \rightarrow \infty$ ,  $i \rightarrow \infty$ , for every  $\Theta \in \mathcal{U}$ .

For proof, see Appendix A.3. Note that every quadratic constraint  $u^T Q u + 2q^T u + r \leq 0$  which is valid on  $U$  induces a *linear constraint*  $\text{Tr} \left( \left[ \begin{array}{c|c} Q & q \\ \hline q^T & r \end{array} \right] Z \right) \leq 0$  which is valid for all matrices  $Z(u)$ ,  $u \in U$ , and thus can be incorporated into the description of  $\mathcal{Z}$ .

**Special case.** In the situation of Proposition 5.1, let  $u$  vary in a convex compact set  $U$ . In this case, the simplest way to define  $\mathcal{Z}$  such that  $Z(u) \in \mathcal{Z}$  for all  $u \in U$  is to set

$$\mathcal{Z} = \left\{ W = \left[ \begin{array}{c|c} V & u \\ \hline u^T & 1 \end{array} \right] : W \succeq 0, u \in U \right\}.$$

Let us compute the function  $\Phi(h, 0; \Theta)$ . Setting  $A = [\bar{A}, a]$ , where  $a$  is the last column of  $A$ , direct computation yields

$$\begin{aligned} \Phi(h, 0; \Theta) &= \frac{1}{2} \phi_{\mathcal{Z}} \left( \left[ \begin{array}{c|c} \bar{A}^T h & \\ \hline h^T \bar{A} & 2a^T h + h^T \Theta_* h \end{array} \right] \right) = \frac{1}{2} \max_{V, u} \left\{ 2u^T \bar{A}^T h + 2a^T h + h^T \Theta_* h : \left[ \begin{array}{c|c} V & u \\ \hline u^T & 1 \end{array} \right] \succeq 0, u \in U \right\} \\ &= \max_{u \in U} \left\{ h^T A u + \frac{1}{2} h^T \Theta_* h \right\}. \end{aligned}$$

Now imagine that we are given two collections  $(A_\chi, U_\chi, \mathcal{U}_\chi)$ ,  $\chi = 1, 2$ , of the  $(A, U, \mathcal{U})$ -data, with the same number of rows in  $A_1$  and  $A_2$  and have associated with  $\mathcal{U}_\chi$   $\succeq$ -upper bounds  $\Theta_{*,\chi}$  on matrices  $\Theta \in \mathcal{U}_\chi$ . We want to use Proposition 5.1 to build an affine detector capable to decide on the hypotheses  $H_1$  and  $H_2$  on the distribution of observation  $\omega$ , with  $H_\chi$  stating that this observation is  $\mathcal{N}(A_\chi[u; 1], \Theta)$  with some  $u \in U_\chi$  and  $\Theta \in \mathcal{U}_\chi$ . To this end, we have to solve the convex-concave saddle point problem

$$\mathcal{SV} = \min_h \max_{\substack{\Theta_1 \in \mathcal{U}_1, \\ \Theta_2 \in \mathcal{U}_2}} \frac{1}{2} [\Phi_1(-h, 0; \Theta_1) + \Phi_2(h, 0; \Theta_2)],$$

where  $\Phi_1, \Phi_2$  are the functions associated, as explained in Proposition 5.1, with the first, respectively, with the second collection of the  $(A, U, \mathcal{U})$ -data. In view of the above computation, this boils down to the necessity to solve the convex minimization problem

$$\mathcal{SV} = \min_h \left\{ \max_{\substack{u_1 \in U_1, \\ u_2 \in U_2}} \frac{1}{2} \left[ \frac{1}{2} h^T \Theta_{*,1} h + \frac{1}{2} h^T \Theta_{*,2} h + h^T [A_2[u_2; 1] - A_1[u_1; 1]] \right] \right\}.$$

An optimal solution  $h_*$  to this problem induces the affine detector

$$\phi_*(\omega) = h_*^T \omega + a, \quad a = \frac{1}{2} \left[ \frac{1}{2} h_*^T [\Theta_{*,1} - \Theta_{*,2}] h_* + \max_{u_1 \in U_1} [-h_*^T A_1[u_1; 1]] - \max_{u_2 \in U_2} h_*^T A_2[u_2; 1] \right],$$

and the risk of this detector on the pair of families  $\mathcal{G}_1, \mathcal{G}_2$  of Gaussian distributions in question is  $\exp\{\mathcal{SV}\}$ .

On the other hand, we could build affine detector for the families  $\mathcal{G}_1, \mathcal{G}_2$  by the machinery from section 3.2.3, that is, by solving convex-concave saddle point problem

$$\overline{\mathcal{SV}} = \min_h \max_{\substack{u_1 \in U_1, \Theta_1 \in \mathcal{U}_1, \\ u_2 \in U_2, \Theta_2 \in \mathcal{U}_2}} \frac{1}{2} \left[ -h^T A_1[u_1; 1] + h^T A_2[u_2; 1] + \frac{1}{2} h^T \Theta_1 h + \frac{1}{2} h^T \Theta_2 h \right];$$

the risk of the resulting affine detector on  $\mathcal{G}_1, \mathcal{G}_2$  is  $\exp\{\overline{\mathcal{SV}}\}$ . Now assume that

(!)  $\mathcal{U}_\chi$ ,  $\chi = 1, 2$ , have  $\succeq$ -maximal elements, and these elements are selected as  $\Theta_{*,\chi}$ .

$\rho$	$\sigma_1$	$\sigma_2$	unrestricted $H$ and $h$	$H = 0$	$h = 0$
0.5	2	2	0.31	0.31	1.00
0.5	1	4	0.24	0.39	0.62
0.01	1	4	0.41	1.00	0.41

**Table 1:** Risk of quadratic detector

In this case the above computation says that  $\mathcal{SV}$  and  $\overline{\mathcal{SV}}$  are the minimal values of identically equal to each other functions and thus are equal to each other. Thus, in the case of (!) the machinery of Proposition 5.1 produces a quadratic detector which can be only better, in terms of risk, than the affine detector yielded by Proposition 3.3.<sup>8</sup>

**Numerical illustration.** To get an impression of the performance of quadratic detectors as compared to affine ones in the case of (!), we present here the results of experiment where  $U_1 = U_1^\rho = \{u \in \mathbf{R}^{12} : u_i \geq \rho, 1 \leq i \leq 12\}$ ,  $U_2 = U_2^\rho = -U_1^\rho$ ,  $A_1 = A_2 \in \mathbf{R}^{8 \times 12}$ , and  $\mathcal{U}_\chi = \{\Theta_{*,\chi} = \sigma_\chi^2 I_8\}$  are singletons. The risks of affine, quadratic and “purely quadratic” (with  $h$  set to 0) detectors on the pair  $\mathcal{G}_1, \mathcal{G}_2$  of families of Gaussian distributions, with  $\mathcal{G}_\chi = \{\mathcal{N}(\theta, \Theta_{*,\chi}) : \theta \in A_\chi U_\chi^\rho\}$ , are given in Table 1.

We see that

- when deciding on families of Gaussian distributions with common covariance matrix and expectations varying in associated with the families convex sets, passing from affine detectors described by Proposition 3.3 to quadratic detectors, does not affect the risk (first row in the table). This is a general fact: by the results of [13], in the situation in question affine detectors are optimal in terms of risk among *all possible* detectors.
- when deciding on families of Gaussian distributions in the case where distributions from different families can have close expectations (third row in the table), affine detectors are useless, while the quadratic ones are not, provided that  $\Theta_{*,1}$  differs from  $\Theta_{*,2}$ . This is how it should be – we are in the case where the first moments of the distribution of the observation bear no definitive information on the family this distribution belongs to, which makes affine detectors useless. In contrast, quadratic detectors are able to utilize information (valuable when  $\Theta_{*,1} \neq \Theta_{*,2}$ ) “stored” in the second moments of the observation.
- “in general” (second row in the table), both affine and purely quadratic components in a quadratic detector are useful; suppressing one of them can increase significantly the attainable risk.

### 5.1.2 Quadratic lifting: Bounded observations

It is convenient to represent a “quadratically lifted observation”  $(\zeta, \zeta^T)$  by the matrix

$$Z(\zeta) = \left[ \begin{array}{c|c} \zeta \zeta^T & \zeta \\ \hline \zeta^T & 1 \end{array} \right] \in \mathbf{S}^{d+1}.$$

<sup>8</sup>This seems to be tautology – there are more quadratic detectors than affine ones, so that the best risk achievable with quadratic detectors can be only smaller than the best risk achievable with affine detectors. The point, however, is that Proposition 5.1 does not guarantee building the best, in terms of its risk, quadratic detector, it deals with “computationally tractable approximation” of this problem. As a result, the quadratic detector constructed in the latter proposition can, in principle, be worse than the affine detector yielded by Proposition 3.3.

Assume that all distributions from  $\mathcal{P}$  are supported on the solution set  $\mathcal{X}$  of a system of quadratic constraints

$$f_\ell(\zeta) := \zeta^T A_\ell \zeta + 2a_\ell^T \zeta + \alpha_\ell \leq 0, \quad 1 \leq \ell \leq L,$$

where  $A_\ell \in \mathbf{S}^d$  are such that  $\sum_\ell \bar{\lambda}_\ell A_\ell \succ 0$  for properly selected  $\bar{\lambda}_\ell \geq 0$ ; as a consequence,  $\mathcal{X}$  is bounded (since  $\bar{f}(\zeta) := \sum_{\ell=1}^L \bar{\lambda}_\ell f_\ell(\zeta)$  is a strongly convex quadratic form which is  $\leq 0$  on  $\mathcal{X}$ ). Setting

$$Q_0 = \begin{bmatrix} \text{---} & \text{---} \\ \text{---} & 1 \end{bmatrix}, Q_\ell = \begin{bmatrix} A_\ell & a_\ell \\ a_\ell^T & \alpha_\ell \end{bmatrix}, \quad 1 \leq \ell \leq L,$$

observe that the distribution  $P^+$  of  $Z(\zeta)$  induced by a distribution  $P \in \mathcal{P}$  is supported on the closed convex set

$$\mathcal{X}^+ = \{Z \in \mathbf{S}^{d+1} : Z \succeq 0, \text{Tr}(Q_0 Z) = 1, \text{Tr}(Q_\ell Z) \leq 0, \ell = 1, \dots, L\}$$

which is bounded<sup>9</sup>. The support function of this set is

$$\phi_{X^+}(H) = \max_{Z \in \mathcal{X}^+} \text{Tr}(HZ) = \max_Z \left\{ \text{Tr}(HZ) : \begin{cases} Z \succeq 0, Z_{d+1,d+1} = 1 \\ \text{Tr}(Q_\ell Z) \leq 0, 1 \leq \ell \leq L \end{cases} \right\}$$

Recalling Example 4 in section 2.2 and section 2.3.5, we arrive at the regular data

$$\mathcal{H} = \mathbf{S}^{d+1}, \mathcal{M} = \mathcal{X}^+, \Phi(h, \mu) = \inf_{g \in \mathbf{S}^{d+1}} \left\{ \text{Tr}((h - g)\mu) + \frac{1}{8}[\phi_{X^+}(h - g) + \phi_{X^+}(g - h)]^2 + \phi_{X^+}(g) \right\}$$

such that

$$\forall P \in \mathcal{P} : P^+ \in \mathcal{S}[\mathbf{S}^{d+1}, \{e[P^+]\}, \Phi(\cdot, e[P^+])],$$

where  $e[P^+]$  is the expectation of  $P^+$ , and therefore

$$\mathcal{P}^+ = \{P^+ : P \in \mathcal{P}\} \subset \mathcal{S}[\mathbf{S}^{d+1}, \mathcal{M}, \Phi].$$

## References

- [1] L. Birgé. *Approximation dans les espaces métriques et théorie de l'estimation: inégalités de Cràmer-Chernoff et théorie asymptotique des tests*. PhD thesis, Université Paris VII, 1980.
- [2] L. Birgé. Vitesses maximales de décroissance des erreurs et tests optimaux associés. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 55(3):261–273, 1981.
- [3] L. Birgé. Sur un théorème de minimax et son application aux tests. *Probab. Math. Stat.*, 3:259–282, 1982.
- [4] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 65(2):181–237, 1983.
- [5] L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. In *Annales de l'Institut Henri Poincaré (B) Probability and Statistics*, volume 42, pages 273–325. Elsevier, 2006.
- [6] L. Birgé. Robust tests for model selection. In M. Banerjee, F. Bunea, J. Huang, V. Koltchinskii, and M. Maathuis, editors, *From Probability to Statistics and Back: High-Dimensional Models and Processes – A Festschrift in Honor of Jon A. Wellner*, pages 47–64. Institute of Mathematical Statistics, 2013.

---

<sup>9</sup>indeed, for large  $\lambda_0 > 0$ , the matrix  $Q = \lambda_0 Q_0 + \sum_{\ell=1}^L \bar{\lambda}_\ell Q_\ell$  is positive definite, and we conclude that  $\mathcal{X}^+$  is contained in the bounded set  $\{Z \in \mathbf{S}^{d+1} : Z \succeq 0, \text{Tr}(QZ) \leq \lambda_0\}$ .

- [7] M. Burnashev. On the minimax detection of an imperfectly known signal in a white noise background. *Theory Probab. Appl.*, 24:107–119, 1979.
- [8] M. Burnashev. Discrimination of hypotheses for gaussian measures and a geometric characterization of the gaussian distribution. *Math. Notes*, 32:757–761, 1982.
- [9] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952.
- [10] D. Donoho and R. Liu. Geometrizing rate of convergence I. Technical report, Tech. Report 137a, Dept. of Statist., University of California, Berkeley, 1987.
- [11] D. L. Donoho and R. C. Liu. Geometrizing rates of convergence, II. *The Annals of Statistics*, pages 633–667, 1991.
- [12] A. Goldenshluger. A universal procedure for aggregating estimators. *The Annals of Statistics*, pages 542–568, 2009.
- [13] A. Goldenshluger, A. Juditsky, and A. Nemirovski. Hypothesis testing by convex optimization. *Electronic Journal of Statistics*, 9(2):1645–1712, 2015.
- [14] J.-B. Hiriart-Urruty and C. Lemarechal. Convex analysis and minimization algorithms i: Fundamentals (grundlehren der mathematischen wissenschaften). 1993.
- [15] Y. Ingster and I. A. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*, volume 169 of *Lecture Notes in Statistics*. Springer, 2002.
- [16] C. Kraft. Some conditions for consistency and uniform consistency of statistical procedures. *Univ. of California Publ. Statist.*, 2:493–507, 1955.
- [17] L. Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

## A Proofs

### A.1 Proof of Proposition 2.1

All we need is to verify (5) and to check that the right hand side function in this relation is convex. The latter is evident, since  $\phi_X(h) + \phi_X(-h) \geq 2\phi_X(0) = 0$  and  $\phi_X(h) + \phi_X(-h)$  is convex. To verify (5), let us fix  $P \in \mathcal{P}[X]$  and  $h \in \mathbf{R}^d$  and set

$$\nu = h^T e[P],$$

so that  $\nu$  is the expectation of  $h^T \omega$  with  $\omega \sim P$ . Note that  $-\phi_X(-h) \leq \nu \leq \phi_X(h)$ , so that (5) definitely holds true when  $\phi_X(h) + \phi_X(-h) = 0$ . Now let

$$\eta := \frac{1}{2} [\phi_X(h) + \phi_X(-h)] > 0,$$

and let

$$a = \frac{1}{2} [\phi_X(h) - \phi_X(-h)], \quad \beta = (\nu - a)/\eta.$$

Denoting by  $P_h$  the distribution of  $h^T \omega$  induced by the distribution  $P$  of  $\omega$  and noting that this distribution is supported on  $[-\phi_X(-h), \phi_X(h)] = [a - \eta, a + \eta]$  and has expectation  $\nu$ , we get

$$\beta \in [-1, 1]$$

and

$$\gamma := \int \exp\{h^T \omega\} P(d\omega) = \int_{a-\eta}^{a+\eta} [e^s - \lambda(s - \nu)] P_h(ds)$$

for all  $\lambda \in \mathbf{R}$ . Hence,

$$\begin{aligned} \ln(\gamma) &\leq \inf_{\lambda} \ln \left( \max_{a-\eta \leq s \leq a+\eta} [e^s - \lambda(s - \nu)] \right) = a + \inf_{\rho} \ln \left( \max_{-\eta \leq t \leq \eta} [e^t - \rho(t - [\nu - a])] \right) \\ &= a + \inf_{\rho} \ln \left( \max_{-\eta \leq t \leq \eta} [e^t - \rho(t - \eta\beta)] \right) \leq a + \ln \left( \max_{-\eta \leq t \leq \eta} [e^t - \bar{\rho}(t - \eta\beta)] \right) \end{aligned}$$

with  $\bar{\rho} = (2\eta)^{-1}(e^\eta - e^{-\eta})$ . The function  $g(t) = e^t - \bar{\rho}(t - \eta\beta)$  is convex on  $[-\eta, \eta]$ , and

$$g(-\eta) = g(\eta) = \cosh(\eta) + \beta \sinh(\eta),$$

which combines with the above computation to yield the relation

$$\ln(\gamma) \leq a + \ln(\cosh(\eta) + \beta \sinh(\eta)), \quad (40)$$

and all we need to verify is that

$$\forall(\eta > 0, \beta \in [-1, 1]) : \beta\eta + \frac{1}{2}\eta^2 - \ln(\cosh(\eta) + \beta \sinh(\eta)) \geq 0. \quad (41)$$

Indeed, if (41) holds true (40) implies that

$$\ln(\gamma) \leq a + \beta\eta + \frac{1}{2}\eta^2 = \nu + \frac{1}{2}\eta^2,$$

which, recalling what  $\gamma$ ,  $\nu$  and  $\eta$  are, is exactly what we want to prove.

Verification of (41) is as follows. The left hand side in (41) is convex in  $\beta$  for  $\beta > -\frac{\cosh(\eta)}{\sinh(\eta)}$  containing, due to  $\eta > 0$ , the range of  $\beta$  in (41). Furthermore, the minimum of the left hand side of (41) over  $\beta > -\coth(\eta)$  is attained when  $\beta = \frac{\sinh(\eta) - \eta \cosh(\eta)}{\eta \sinh(\eta)}$  and is equal to

$$r(\eta) = \frac{1}{2}\eta^2 + 1 - \eta \coth(\eta) - \ln(\sinh(\eta)/\eta).$$

All we need to prove is that the latter quantity is nonnegative whenever  $\eta > 0$ . We have

$$r'(\eta) = \eta - \coth(\eta) - \eta(1 - \coth^2(\eta)) - \coth(\eta) + \eta^{-1} = (\eta \coth(\eta) - 1)^2 \eta^{-1} \geq 0,$$

and since  $r(+0) = 0$ , we get  $r(\eta) \geq 0$  when  $\eta > 0$ .  $\square$

## A.2 Proof of Proposition 4.1

Let  $\mathbf{D}_\ell$  denote the collection of regular data processed by  $\mathcal{A}_K^{\ell\delta_\ell}$ . Let, further, the  $K$ -repeated random observation  $\omega^K$ , the probability distribution  $\bar{P}$ , index  $\bar{i} \leq I$  and vector  $\bar{\mu} \in \mathcal{M}_{\bar{i}}$  be as in the premise of the proposition, let  $\bar{g} = G\bar{\mu}$ , and let  $\ell_* \leq L$  be such that  $\|\bar{g} - g_{\ell_*}\|_2 \leq \|\bar{g} - g_\ell\|_2$ ,  $\ell = 1, \dots, L$ . Finally, let  $\mathcal{E}$  be the event “for every  $\ell = 1, \dots, L$  such that  $\bar{P}$  obeys one or more of the hypotheses  $H_{\mathcal{S}i}[\mathcal{D}]$ ,  $\mathcal{D} \in \mathbf{D}_\ell$ , processed by procedure  $\mathcal{A}_K^{\ell\delta_\ell}$ , this procedure correctly recovers the color of these hypotheses.” By construction and due to the union bound, the  $\bar{P}^K$ -probability of  $\mathcal{E}$  is at least  $1 - \epsilon$ . It follows that all we need to verify the claim of the proposition is to show that when  $\omega^K \in \mathcal{E}$ , relation (33) takes place. Thus, let us fix  $\omega^K \in \mathcal{E}$ .

Observe, first, that  $\bar{g} \in V_{\ell_*}$ , whence  $\bar{\mu} \in W_{\ell_*}^{\bar{i}}$ . Thus, when running  $\mathcal{A}_K^{\ell_*\delta_{\ell_*}}$ ,  $\bar{P}^K$  obeys a red one among the hypotheses  $H_{\mathcal{S}i}[\mathcal{D}]$ ,  $\mathcal{D} \in \mathbf{D}_{\ell_*}$  processed by  $\mathcal{A}_K^{\ell_*\delta_{\ell_*}}$ , and since we are in the case of  $\omega^K \in \mathcal{E}$ ,

$g_{\ell_*}$  gets a color, namely, color “red.” By construction of our aggregation procedure, its output can be either  $g_{\ell_*}$  – and in this case (33) clearly holds true, or another vector, let it be denoted  $g_{\ell_+}$  ( $\ell_+ \neq \ell_*$ ), which was also assigned red color. We claim that the vector  $\bar{g} = G\bar{\mu}$  satisfies the relation

$$u_{\ell_+ \ell_*}^T \bar{g} < v_{\ell_+ \ell_*} + \delta_{\ell_+}. \quad (42)$$

Indeed, otherwise we have  $\bar{\mu} \in W_{\ell_+ \ell_*}^{\bar{\delta}_{\ell_+}}$ , meaning that  $\bar{P}^K$  obeys a hypothesis  $H_{S_i}[\mathcal{D}]$  processed when running  $\mathcal{A}_K^{\ell_+ \delta_{\ell_+}}$  (i.e., with  $\mathcal{D} \in \mathbf{D}_{\ell_+}$ ), and this hypothesis is blue. Since we are in the case of  $\mathcal{E}$ , this implies that the color inferred by  $\mathcal{A}_K^{\ell_+ \delta_{\ell_+}}$  is “blue,” which is a desired contradiction.

Now we are nearly done: indeed,  $g_{\ell_*}$  is the  $\|\cdot\|_2$  closest to  $\bar{g}$  point among  $g_1, \dots, g_L$ , implying that

$$u_{\ell_+ \ell_*}^T \bar{g} \geq v_{\ell_+ \ell_*}. \quad (43)$$

Recalling what  $u_{\ell_* \ell_+}$  and  $v_{\ell_* \ell_+}$  are, the relations (42) and (43) tell us the following story about the points  $\bar{g}$ ,  $g_* := g_{\ell_*}$ ,  $g_+ := g_{\ell_+}$  and the hyperplane  $H = \{g \in \mathbf{R}^m : u_{\ell_* \ell_+} g = v_{\ell_* \ell_+}\}$ :  $g_*$  and  $g_+$  are symmetric to each other w.r.t.  $H$ , and  $\bar{g}$  is at most at the distance  $\delta := \delta_{\ell_+}$  of  $H$ . An immediate observation is that in this case

$$\|g_+ - \bar{g}\|_2 \leq \|g_* - \bar{g}\|_2 + 2\delta, \quad (44)$$

and we arrive at (33).

To justify (44) note that by shift and rotation we can reduce the situation to the one when  $g_*, g_+, \bar{g}$  belong to the linear span of the first two basic orhts with the first two coordinates of these three vectors being, respectively,  $[-r; 0]$  (for  $g_*$ ),  $[r; 0]$  (for  $g_+$ ) and  $[d; h]$  (for  $\bar{g}$ ), with  $|d| \leq \delta$ . Hence

$$\|g_+ - \bar{g}\|_2 - \|g_* - \bar{g}\|_2 = \frac{[(r-d)^2 + h^2] - [(r+d)^2 + h^2]}{\|g_+ - \bar{g}\|_2 + \|g_* - \bar{g}\|_2} = \frac{-4rd}{\|g_+ - \bar{g}\|_2 + \|g_* - \bar{g}\|_2} \leq \frac{4r\delta}{\|g_+ - g_*\|_2} = 2\delta,$$

as claimed.  $\square$

### A.3 Proof of Proposition 5.1

1<sup>0</sup>. For any  $u \in \mathbf{R}^m$ ,  $h \in \mathbf{R}^d$ ,  $\Theta \in \mathbf{S}_+^d$  and  $H \in \mathbf{S}^d$  such that  $-I \prec \Theta^{1/2} H \Theta^{1/2} \prec I$  we have

$$\begin{aligned} \Psi(h, H; u, \Theta) &:= \ln(\mathbf{E}_{\zeta \sim \mathcal{N}(\mathcal{A}(u), \Theta)} \{ \exp\{h^T \zeta + \tfrac{1}{2} \zeta^T H \zeta\} \}) \\ &= \ln(\mathbf{E}_{\xi \sim \mathcal{N}(0, I)} \{ \exp\{h^T [\mathcal{A}(u) + \Theta^{1/2} \xi] + \tfrac{1}{2} [\mathcal{A}(u) + \Theta^{1/2} \xi]^T H [\mathcal{A}(u) + \Theta^{1/2} \xi]\} \}) \\ &= -\tfrac{1}{2} \ln \text{Det}(I - \Theta^{1/2} H \Theta^{1/2}) \\ &\quad + h^T \mathcal{A}(u) + \tfrac{1}{2} \mathcal{A}(u)^T H \mathcal{A}(u) + \tfrac{1}{2} [H \mathcal{A}(u) + h]^T \Theta^{1/2} [I - \Theta^{1/2} H \Theta^{1/2}]^{-1} \Theta^{1/2} [H \mathcal{A}(u) + h] \\ &= -\tfrac{1}{2} \ln \text{Det}(I - \Theta^{1/2} H \Theta^{1/2}) + \tfrac{1}{2} [u; 1]^T [bh^T A + A^T h b^T + A^T H A] [u; 1] \\ &\quad + \tfrac{1}{2} [u; 1]^T [B^T [H, h]^T \Theta^{1/2} [I - \Theta^{1/2} H \Theta^{1/2}]^{-1} \Theta^{1/2} [H, h] B] [u; 1] \end{aligned} \quad (45)$$

(because  $h^T \mathcal{A}(u) = [u; 1]^T b h^T A [u; 1] = [u; 1]^T A^T h b^T [u; 1]$  and  $H \mathcal{A}(u) + h = [H, h] B [u; 1]$ ).

Observe that when  $(h, H) \in \mathcal{H}_\gamma$ , we have  $\Theta^{1/2} [I - \Theta^{1/2} H \Theta^{1/2}]^{-1} \Theta^{1/2} = [\Theta^{-1} - H]^{-1} \preceq [\Theta_*^{-1} - H]^{-1}$ , so that (45) implies that for all  $u \in \mathbf{R}^m$ ,  $\Theta \in \mathcal{U}$ , and  $(h, H) \in \mathcal{H}_\gamma$ ,

$$\begin{aligned} \Psi(h, H; u, \Theta) &\leq -\tfrac{1}{2} \ln \text{Det}(I - \Theta^{1/2} H \Theta^{1/2}) \\ &\quad + \tfrac{1}{2} [u; 1]^T \underbrace{[bh^T A + A^T h b^T + A^T H A + B^T [H, h]^T [\Theta_*^{-1} - H]^{-1} [H, h] B]}_{Q[H, h]} [u; 1] \\ &= -\tfrac{1}{2} \ln \text{Det}(I - \Theta^{1/2} H \Theta^{1/2}) + \tfrac{1}{2} \text{Tr}(Q[H, h] Z(u)) \\ &\leq -\tfrac{1}{2} \ln \text{Det}(I - \Theta^{1/2} H \Theta^{1/2}) + \Gamma_{\mathcal{Z}}(H, h) \end{aligned} \quad (46)$$



(we have taken into account that  $Z(u) \in \mathcal{Z}$  when  $u \in U$  (premise of the proposition) and therefore  $\text{Tr}(Q[H, h]Z(u)) \leq \phi_{\mathcal{Z}}(Q[H, h])$ ).

**2<sup>0</sup>.** Now let us upper-bound the function

$$G(h, H; \Theta) = -\frac{1}{2} \ln \text{Det}(I - \Theta^{1/2} H \Theta^{1/2})$$

on the domain  $(h, H) \in \mathcal{H}_\gamma$ ,  $\Theta \in \mathcal{U}$ . For  $(h, H) \in \mathcal{H}_\gamma$  and  $\Theta \in \mathcal{U}$  fixed we have

$$\begin{aligned} \|\Theta^{1/2} H \Theta^{1/2}\| &= \|[\Theta^{1/2} \Theta_*^{-1/2}][\Theta_*^{1/2} H \Theta_*^{1/2}][\Theta^{1/2} \Theta_*^{-1/2}]^T\| \\ &\leq \|\Theta^{1/2} \Theta_*^{-1/2}\|^2 \|\Theta_*^{1/2} H \Theta_*^{1/2}\| \leq \|\Theta_*^{1/2} H \Theta_*^{1/2}\| =: d(H) \leq \gamma \end{aligned} \quad (47)$$

(we have used the fact that  $0 \preceq [\Theta^{1/2} \Theta_*^{-1/2}]^T [\Theta^{1/2} \Theta_*^{-1/2}] \preceq I$  due to  $0 \preceq \Theta \preceq \Theta_*$ , whence  $\|\Theta^{1/2} \Theta_*^{-1/2}\| \leq 1$ ). Denoting by  $\|\cdot\|_F$  the Frobenius norm of a matrix and noting that  $\|AB\|_F \leq \|A\| \|B\|_F$ , computation completely similar to the one in (47) yields

$$\|\Theta^{1/2} H \Theta^{1/2}\|_F \leq \|\Theta_*^{1/2} H \Theta_*^{1/2}\|_F =: D(H). \quad (48)$$

Besides this, setting  $F(X) = -\ln \text{Det}(X) : \text{int } \mathbf{S}_+^d \rightarrow \mathbf{R}$  and equipping  $\mathbf{S}^d$  with the Frobenius inner product, we have  $\nabla F(X) = -X^{-1}$ , so that with  $R_0 = \Theta_*^{1/2} H \Theta_*^{1/2}$ ,  $R_1 = \Theta^{1/2} H \Theta^{1/2}$ , and  $\Delta = R_1 - R_0$ , we have for properly selected  $\lambda \in (0, 1)$  and  $R_\lambda = \lambda R_0 + (1 - \lambda) R_1$ :

$$\begin{aligned} F(I - R_1) &= F(I - R_0 - \Delta) = F(I - R_0) + \langle \nabla F(I - R_\lambda), -\Delta \rangle = F(I - R_0) + \langle (I - R_\lambda)^{-1}, \Delta \rangle \\ &= F(I - R_0) + \langle I, \Delta \rangle + \langle (I - R_\lambda)^{-1} - I, \Delta \rangle. \end{aligned}$$

We conclude that

$$F(I - R_1) \leq F(I - R_0) + \text{Tr}(\Delta) + \|I - (I - R_\lambda)^{-1}\|_F \|\Delta\|_F. \quad (49)$$

Denoting by  $\mu_i$  the eigenvalues of  $R_\lambda$  and noting that  $\|R_\lambda\| \leq \max[\|R_0\|, \|R_1\|] = d(H) \leq \gamma$  (see (47)), we have  $|\mu_i| \leq \gamma$ , and therefore eigenvalues  $\nu_i = 1 - \frac{1}{1 - \mu_i} = -\frac{\mu_i}{1 - \mu_i}$  of  $I - (I - R_\lambda)^{-1}$  satisfy  $|\nu_i| \leq |\mu_i|/(1 - \mu_i) \leq |\mu_i|/(1 - \gamma)$ , whence

$$\|I - (I - R_\lambda)^{-1}\|_F \leq \|R_\lambda\|_F / (1 - \gamma).$$

Noting that  $\|R_\lambda\|_F \leq \max[\|R_0\|_F, \|R_1\|_F] \leq D(H)$ , see (48), we conclude that  $\|I - (I - R_\lambda)^{-1}\|_F \leq D(H)/(1 - \gamma)$ , so that (49) yields

$$F(I - R_1) \leq F(I - R_0) + \text{Tr}(\Delta) + D(H) \|\Delta\|_F / (1 - \gamma). \quad (50)$$

Further, by (37) the matrix  $D = \Theta^{1/2} \Theta_*^{-1/2} - I$  satisfies  $\|D\| \leq \delta$ , whence

$$\Delta = \underbrace{\Theta^{1/2} H \Theta^{1/2}}_{R_1} - \underbrace{\Theta_*^{1/2} H \Theta_*^{1/2}}_{R_0} = (I + D) R_0 (I + D^T) - R_0 = D R_0 + R_0 D^T + D R_0 D^T.$$

Consequently,

$$\|\Delta\|_F \leq \|D R_0\|_F + \|R_0 D^T\|_F + \|D R_0 D^T\|_F \leq [2\|D\| + \|D\|^2] \|R_0\|_F \leq \delta(2 + \delta) \|R_0\|_F = \delta(2 + \delta) D(H).$$

This combines with (50) and the relation

$$\text{Tr}(\Delta) = \text{Tr}(\Theta^{1/2} H \Theta^{1/2} - \Theta_*^{1/2} H \Theta_*^{1/2}) = \text{Tr}([\Theta - \Theta_*] H)$$

to yield

$$F(I - R_1) \leq F(I - R_0) + \text{Tr}([\Theta - \Theta_*]H) + \frac{\delta(2 + \delta)}{1 - \gamma} \|\Theta_*^{1/2} H \Theta_*^{1/2}\|_F^2,$$

and we conclude that for all  $(h, H) \in \mathcal{H}_\gamma$  and  $\Theta \in \mathcal{U}$ ,

$$\begin{aligned} G(h, H; \Theta) &= \frac{1}{2} F(I - R_1) \\ &\leq \widehat{G}(h, H; \Theta) := -\frac{1}{2} \ln \text{Det}(I - \Theta_*^{1/2} H \Theta_*^{1/2}) + \frac{1}{2} \text{Tr}([\Theta - \Theta_*]H) + \frac{\delta(2 + \delta)}{2(1 - \gamma)} \|\Theta_*^{1/2} H \Theta_*^{1/2}\|_F^2. \end{aligned} \quad (51)$$

Note that  $\widehat{G}(h, H; \Theta)$  clearly is convex-concave and continuous on  $\mathcal{H} \times \mathcal{M} = \mathcal{H}_\gamma \times \mathcal{U}$ .

**3<sup>0</sup>.** Combining (51), (46), (38) and the origin of  $\Psi$ , see (45), we arrive at

$$\forall ((u, \Theta) \in U \times \mathcal{U}, (h, H) \in \mathcal{H}_\gamma = \mathcal{H}) : \ln(\mathbf{E}_{\zeta \sim \mathcal{N}(u, \Theta)} \{ \exp\{h^T \zeta + \frac{1}{2} \zeta^T H \zeta\} \}) \leq \Phi(h, H; \Theta).$$

This is all we need, up to verification of the claim that  $\mathcal{H}, \mathcal{M}, \Phi$  is regular data, which boils down to checking that  $\Phi : \mathcal{H} \times \mathcal{M} \rightarrow \mathbf{R}$  is convex-concave and continuous. The latter check, recalling that  $\widehat{G}(h, H; \Theta) : \mathcal{H} \times \mathcal{M}$  indeed is convex-concave and continuous, reduces to verifying that  $\Gamma(h, H)$  is convex and continuous on  $\mathcal{H}_\gamma$ . Recalling that  $\mathcal{Z}$  is nonempty compact set, the function  $\phi_{\mathcal{Z}}(\cdot) : \mathbf{S}^{d+1}$  is continuous, implying the continuity of  $\Gamma(h, H) = \frac{1}{2} \phi_{\mathcal{Z}}(Q[H, h])$  on  $\mathcal{H} = \mathcal{H}_\gamma$  ( $Q[H, h]$  is defined in (46)). To prove convexity of  $\Gamma$ , note that  $\mathcal{Z}$  is contained in  $\mathbf{S}_+^{m+1}$ , implying that  $\phi_{\mathcal{Z}}(\cdot)$  is convex and  $\succeq$ -monotone. On the other hand, by Schur Complement Lemma, we have

$$\begin{aligned} S &:= \{(h, H, G) : G \succeq Q[H, h], (h, H) \in \mathcal{H}_\gamma\} \\ &= \left\{ (h, H, G) : \left[ \begin{array}{c|c} G - [bh^T A + A^T h B + A^T H A] & B^T [H, h]^T \\ \hline [H, h] B & \Theta_*^{-1} - H \end{array} \right] \succeq 0, (h, H) \in \mathcal{H}_\gamma \right\}, \end{aligned}$$

implying that  $S$  is convex. Since  $\Phi_{\mathcal{Z}}(\cdot)$  is  $\succeq$ -monotone, we have

$$\{(h, H, \tau) : (h, H) \in \mathcal{H}_\gamma, \tau \geq \Gamma(h, H)\} = \{(h, H, \tau) : \exists G : G \succeq Q[H, h], 2\tau \geq \phi_{\mathcal{Z}}(G), (h, H) \in \mathcal{H}_\gamma\},$$

and we see that the epigraph of  $\Gamma$  is convex (since the set  $S$  and the epigraph of  $\phi_{\mathcal{Z}}$  are so), as claimed.

**4<sup>0</sup>.** It remains to prove that  $\Phi$  is coercive in  $H, h$ . Let  $\Theta \in \mathcal{U}$  and  $(h_i, H_i) \in \mathcal{H}_\gamma$  with  $\|(h_i, H_i)\| \rightarrow \infty$  as  $i \rightarrow \infty$ , and let us prove that  $\Phi(h_i, H_i; \Theta) \rightarrow \infty$ . Looking at the expression for  $\Phi(h_i, H_i; \Theta)$ , it is immediately seen that all terms in this expression, except for the terms coming from  $\phi_{\mathcal{Z}}(\cdot)$ , remain bounded as  $i$  grows, so that all we need to verify is that the  $\phi_{\mathcal{Z}}(\cdot)$ -term goes to  $\infty$  as  $i \rightarrow \infty$ . Observe that  $H_i$  are uniformly bounded due to  $(h_i, H_i) \in \mathcal{H}_\gamma$ , implying that  $\|h_i\|_2 \rightarrow \infty$  as  $i \rightarrow \infty$ . Denoting by  $e$  the last basic orth of  $\mathbf{R}^{d+1}$  and by  $b$ , as before, the last basic orth of  $\mathbf{R}^{m+1}$ , note that, by construction,  $B^T e = b$ . Now let  $W \in \mathcal{Z}$ , so that  $W_{m+1, m+1} = 1$ . Taking into account that the matrices  $[\Theta_*^{-1} - H_i]^{-1}$  satisfy  $\alpha I_d \preceq [\Theta_*^{-1} - H_i]^{-1} \preceq \beta I_d$  for some positive  $\alpha, \beta$  due to  $H_i \in \mathcal{H}_\gamma$ , observe that

$$\underbrace{\left[ \left[ \begin{array}{c|c} H_i & h_i \\ \hline h_i^T & \end{array} \right] + [H_i, h_i]^T [\Theta_*^{-1} - H_i]^{-1} [H_i, h_i] \right]}_{Q_i} = \underbrace{[h_i^T [\Theta_*^{-1} - H_i]^{-1} h_i]}_{\alpha_i \|h_i\|_2^2} e e^T + R_i,$$

where  $\alpha_i \geq \alpha > 0$  and  $\|R_i\|_F \leq C(1 + \|h_i\|_2)$ . As a result,

$$\begin{aligned} \phi_{\mathcal{Z}}(B^T Q_i B) &\geq \text{Tr}(W B^T Q_i B) = \text{Tr}(W B^T [\alpha_i \|h_i\|_2^2 e e^T + R_i] B) \\ &= \alpha_i \|h_i\|_2^2 \underbrace{\text{Tr}(W b b^T)}_{=W_{m+1, m+1}=1} - \|B W B^T\|_F \|R_i\|_F \geq \alpha \|h_i\|_2^2 - C(1 + \|h_i\|_2) \|B W B^T\|_F, \end{aligned}$$

and the concluding quantity tends to  $\infty$  as  $i \rightarrow \infty$  due to  $\|h_i\|_2 \rightarrow \infty, i \rightarrow \infty$ .  $\square$